AD_____

Award Number: DAMD17-01-1-0376

TITLE: Investigating the Mechanisms of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

PRINCIPAL INVESTIGATOR: Albert R. Cunningham, Ph.D.

CONTRACTING ORGANIZATION: Louisiana State University
Baton Rouge, LA 70803-3701

REPORT DATE: August 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-08-2005	Annual	15 Jul 2004 – 14 Jul 2005

4. TITLE AND SUBTITLE
Investigating the Mechanisms of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-01-1-0376

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Albert R. Cunningham, Ph.D.

E-mail: arc@lsu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Louisiana State University
Baton Rouge, LA 70803-3701

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project is investigating the potential that environmental chemicals may be involved in the etiology of breast cancer. We hypothesize that specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective is to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the development of breast cancer. With the success of the rat and mouse mammary carcinogen models we have submitted one (accepting pending revisions) and are preparing another publication. We are also pursuing work on a novel SAR approach that is allowing us to ask the specific question of "why do some carcinogens cause cancer in the breast?" which is a significantly different question than posed in historical SAR studies as "why do some chemicals cause cancer?".. Two graduate students have been awarded MS degrees from LSU based on this project, one of whom was funded through this project. We have also successfully used the data and protocols generated from this project to successfully compete for additional funding effectively extending this work from studying "why chemicals cause breast cancer" to "how can new highly specific breast cancer chemotherapies be discovered". Looking forward, I see no obstacles to the successful completion of this project in a timely manner.

15. SUBJECT TERMS
Structure-activity relationship (SAR), computer modeling, mechanisms and etiology of breast cancer, environmental carcinogens

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	62	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Manuscripts:

Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers

A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens

Annual Review August 2005

Investigating the Mechanism of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

DAMD17-01-1-0376

Albert R. Cunningham, Ph.D.

Note on One Year No Cost Extension

As anticipated during our move from the University of Pittsburgh to Louisiana State University, we would require a one year no cost extension. Therefore, this report is an annual report. The final report will be due next year.

Introduction

The well-established breast cancer risk factors may account for only 47% of the breast cancer incidence in the United States. This leaves a considerable portion of breast cancer from undetermined origin. This project is investigating the potential that environmental chemicals and particularly those with estrogenic activity may be involved in the etiology of breast cancer. We hypothesize that specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective is to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the development of breast cancer. The successful completion of this project will provide mechanistic information related to chemical-induced breast cancer as well as structure-activity relationship (SAR) models capable of estimating the likelihood that chemicals with unknown carcinogenic activity may be breast carcinogens.

Body

Software Change

As we noted in our 2004 report, the SAR modeling for this project was originally proposed to be conducted with the MCASE program. However, for multiple reasons we decided to develop our own system. This change did not alter the project and an updated Statement of Work was provided during the past year. I have discussed this matter with Dr. Moore. I am including the updated (but not yet approved) Statement of Work in the appendices.

We are happy to report that the submitted publication to *SAR and QSAR in Environmental Research* was accepted and published in April of 2005. The manuscript described the cat-SAR program in detail. We note the publication was on respiratory sensitizers—not breast carcinogens. The reason for this was 1) it was a small and manageable dataset and 2) a previous MCASE analysis of this data yielded a very good model. As such, this was a suitable dataset on which to develop and test the cat-SAR program. A copy of the publication is included in the appendices.

Specific Aim Accomplishments

The Specific Aims for the project are:

Specific aim 1: Development and validation of SAR models for female breast carcinogens (months 1-12).

a. Identify chemicals tested in female rodents from the Carcinogenic Potency Database and the National Toxicology Program (month 1).
b. Enter chemical structures and potency values into MCASE program (months 2-8).
c. Validate models using 10-fold cross validation (months 9-12)
d. Summarize and interpret models and prepare publication.

As previously discussed, these models have been developed and validated (i.e., a-c) as planned in MCASE and then with the cat-SAR program. We have also updated rodent carcinogenicity models so that all models (mouse and rat, as well as female specific version) have been built on the same datasets and analyzed with the cat-SAR program. We are preparing to publish two manuscripts describing mouse and rat mammary carcinogens.

Female and Mammary Carcinogen Models

We have had success in developing the female and mammary carcinogen models and have devoted a significant effort in "assuring" their appropriateness. Basically, we have now developed several different female and mammary carcinogen models.

1. Rat Mammary Carcinogen Models: In the previous Annual Report we speculated that our original model protocol might not have been optimal. The common approach of most SAR studies entails a comparison of structural features between biologically "active" and "inactive" compounds. Thus, when considering carcinogenesis, the categories are clearly carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis as in the case of breast cancer, we asked the question "what are the appropriate inactive compounds?" Should they be noncarcinogens or carcinogens that are just not carcinogenic to the organ under study? As we proceeded we considered both options. We note again this important aspect of the project was not considered in the original proposal. Moreover, as we investigated this approach, it became evident that this modeling paradigm is to the best of our knowledge, novel.

We developed two separate models for rat mammary carcinogens: The mammary carcinogen - noncarcinogen model and the mammary carcinogen - non-mammary carcinogen model. The details were described in the previous Annual Report. We are happy to report that this new approach to modeling organ-specific carcinogenicity was successful and a manuscript has been accepted in *Chemical Research in Toxicology* pending revisions. The manuscript entitled "A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens" is included in the appendices.

One set of models for this manuscript was developed based on a comparison of rat mammary carcinogens to noncarcinogens (MC-NC) and the second and novel method compared mammary carcinogens to non-mammary carcinogens (MC-NMC). The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88%. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74%. As mentioned, the MC-NMC model was based on a learning set that

contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories and was able to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Based on a structural comparison between this model and one for *Salmonella* mutagens, there was no observed relationship between the two phenomena since both the active and inactive categories contained a high proportion of mutagens. Overall, these findings suggest that the MC-NMC model is identifying structural attributes to address the specific question of "why do some carcinogens cause cancer in the breast?" which is a significantly different question than "why do some chemicals cause cancer?'.

2. Mouse Mammary Carcinogen Models:
A similar group of analyses as listed above for rat carcinogens is being completed for mouse carcinogens. The validation results are shown in Tables 3 and 4. These models are based on 24 mouse mammary carcinogens from the published CPDB target site summary (*15*).

The Second Specific Aim is:
Specific aim 2: Identify chemical and biological attributes of female and/or breast carcinogens to provide evidence to test the hypothesis that xenoestrogens are involved in breast cancer (months 13-36).

a. Compare and identify Structural Feature Overlap Method of female and breast carcinogens to those of other available toxicological SAR models (see Facilities and Equipment for a complete list of available models) (months 13-16).
b. As above using Joint Prevalence Method (months 16-24).
c. Identify the exact features of female and breast carcinogen models that are responsible for predicted similar activities identified above (months 25-26).
d. Conduct QSAR and CoMFA analyses with chemicals containing these structures using biological data from appropriate assays (months 28-36).
e. Conduct metabolism experiments on identified outliers to see whether metabolic activation is required for activity and update models if required (months 28-36).
f. Summarize and interpret data and prepare publications (months 28-36).

We have concluded the migration of a set of approximately 20 MCASE toxicological SAR models to cat-SAR and the models have been validated. This was required for Specific Aims 2a and 2b. We are currently using these models to prepare the Joint Prevalence Method datasets for the 20 endpoints.

During this past year, we also ventured out and developed three estrogen cat-SAR models that will be directly applicable to testing the relationship between estrogenicity and mammary carcinogenicity. A manuscript describing the cat-SAR analysis of 122 chemicals tested for estrogenicity in the E-Screen assay is nearly complete.

We note that in particular SA 3 requires the MCASE module META. Moreover, some of the other sub-SA could be most likely easily accomplished with MCASE. However, as noted in the previous Annual Report we did not have current access to a working copy of MCASE, though

Professor Rosenkranz was working on the issue. Sadly, we must report that Professor Rosenkranz passed away in November of 2004. Due to significant legal issues between himself and Professor Klopman (his co-developer of MCASE), I will not request to have access to MCASE or META.

In retrospect, SA 3, while generally important, is not a practical aim for this project. The reason is that since metabolism is not considered during model development, metabolic products cannot be used to identify outliers. In other words, since the models are built on the parent molecules (i.e., what the animals are dosed with) they do not explicitly consider metabolic products. Thus even if we were to analyze known metabolites of a carcinogen, the models would likely identify the parent structures as relating to carcinogenesis, not the additional metabolism-related moieties.

Key Research Accomplishments

Developed new SAR modeling algorithm called cat-SAR.

Developed predictive and mechanistically insightful SAR models for rat and mouse carcinogens and mammary carcinogens

Development of shareable databases/learning sets of carcinogens, their molecular structures and associated activity values. We are discussing the exact method of "sharing" this information with Dr. Ann Richard of the Environmental Carcinogenesis Division of the U.S. EPA. Dr. Richard has developed and maintains the EPA's DSSTox Public Toxicity Database Network.

Reportable Outcomes-This Period

Graduate Students

We are proud to announce that Shanna Moss and Daniel Consoer both graduated with M.S. degrees from the Department of Environmental Studies at LSU with thesis research based on this project. Specifically, Ms. Moss was supported for her studies through this project and Mr. Consoer, while working on related projects, was funded by LSU.

Ms. Moss's thesis is titled *Identification of 'structural alerts' and associated mechanisms of action of mammary gland carcinogens in female rodents.*

Mr. Consoer's thesis is titled *Evaluation of a Novel Method of Predicting Estrogen Activity of a Group of Structurally Diverse Compounds.*

Manuscripts

Cunningham, A.R., Cunningham, S.L., Consoer, D.M., Moss, S.T. and Karol, M.H. (2005) Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers. SAR QSAR Environ. Res., 16, 273-285.

(Cunningham, A.R., Moss, S.T. and Cunningham, S.L. (accepted pending revisions) A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens. Chem. Res. Toxicol.

Funding Applied for Based on Work Supported by this Award

We note that the below listed proposals all relate to the discovery of novel antibreast cancer therapeutics. Given that the estrogen receptor is involved in the etiology, cure, and prevention of breast cancer, this IDEA Award has allowed us to pursue new avenues of research into drug discovery.

We are happy to report that during the last year we received award notice for an IDEA award titled *A novel approach for the identification of pharmacophores through differential toxicity analysis of estrogen receptor positive and negative cell lines* (PI, $372,542). Our success in achieving this award lies clearly in the work of the project we are now completing. Specifically, the current award has provided us with good set of models on which to understand breast carcinogens and particularly important, has allowed us to develop the basis of a new SAR modeling approach wherein it is possible to differentiate "different" actions of toxicant or carcinogens. As the current IDEA award is now focusing on why particular *carcinogens* are carcinogenic to breast tissue, this new award will focus on understanding and identifying why toxic agents are only specific to certain types of breast tumor cells. The goal is to identify novel and highly selective new chemotherapeutic agents and pharmacophores for their development.

Submitted but not Funded

1. Louisiana Environmental Hazard Survey and Breast Cancer Analysis, The Coypu Foundation (ARC PI, $241,830)
2. Estimating exposure to environmental carcinogens and breast cancer, The Susan G. Komen Breast Cancer Foundation (ARC PI, $249,786)

These first two proposals are based directly on the work of this project wherein we are proposing to apply the predictive capabilities of the models to Geographic Information System (GIS) overlays of breast cancer mortality and morbidity and environmental chemical transport data from Toxic Release Inventory release sites. Only the Komen proposal was returned with review comments and we are pleased with the fact that on this first attempt for funding of this new initiative, the comments were positive and useful. We are hopeful that with further refinement of the proposal, it will become fundable in the near future.

3. Pharmacophore discovery by differential toxicity studies, The Susan G. Komen Breast Cancer Foundation, (Billy Day PI, $249,422)

This proposal was a simpler version of the just discussed IDEA award.

Conclusions

With the success of the rat mammary carcinogen models we are preparing a similar manuscript describing mouse mammary carcinogens. We are also completing work on a general chemical carcinogen manuscript and one describing female-specific carcinogens. Also of importance, we are working on several xenoestrogen and toxicological models that, although not detailed in the project proposal, will be of great importance for studying the receptor- and mutagenesis-based mechanisms of breast carcinogenesis.

To date, after approximately three years of work, we have developed the proposed models set forth in Specific Aim 1 using MCASE. We are now slightly behind schedule due to the time required to develop the cat-SAR program. However, in conjunction with this and other projects in my laboratory, all the required components for Specific Aim 2 are being or will be completed in the coming year. There should be no significant future delays or problems accomplishing the tasks of Specific Aim 2. This is of particular relevance for Specific Aims 2a and 2b that require other toxicological models (e.g., mutagenicity and estrogenicity) on which to compare the female and mammary gland carcinogen models.

Looking forward I see no obstacles to the successful completion of this project in a timely manner which includes the anticipated request for a one-year no cost extension due to our move from the University of Pittsburgh to LSU and the concurrent development of the cat-SAR program.

Finally, in our discussion with friends knowledgeable in SAR and carcinogenesis including the late Professor Rosenkranz, we are very excited about the prospects of this project that lay before us. Specifically, by being able to address the issue of "why chemical carcinogens cause cancer in the breast?" rather than previous SAR-based questions of "why do chemicals cause cancer?" we anticipate taking SAR-based analyses of breast carcinogens to a new level of detail and understanding.

SAR and QSAR in Environmental Research, Vol. 16, No. 3, June 2005, 273–285

Taylor & Francis
Taylor & Francis Group

Development of an information-intensive structure–activity relationship model and its application to human respiratory chemical sensitizers

A. R. CUNNINGHAM†*, S. L. CUNNINGHAM†, D. M. CONSOER†, S. T. MOSS† and M. H. KAROL‡

†Department of Environmental Studies, Louisiana State University, Baton Rouge, LA 70803, USA
‡Department of Environmental and Occupational Health, University of Pittsburgh, Pittsburgh, PA 15261, USA

(Received 20 May 2004; in final form 9 October 2004)

Structure–activity relationship (SAR) models are recognized as powerful tools to predict the toxicologic potential of new or untested chemicals and also provide insight into possible mechanisms of toxicity. Models have been based on physicochemical attributes and structural features of chemicals. We describe herein the development of a new SAR modeling algorithm called cat-SAR that is capable of analyzing and predicting chemical activity from divergent biological response data. The cat-SAR program develops chemical fragment-based SAR models from categorical biological response data (e.g. toxicologically active and inactive compounds). The database selected for model development was a published set of chemicals documented to cause respiratory hypersensitivity in humans. Two models were generated that differed only in that one model included explicate hydrogen containing fragments. The predictive abilities of the models were tested using leave-one-out cross-validation tests. One model had a sensitivity of 0.94 and specificity of 0.87 yielding an overall correct prediction of 91%. The second model had a sensitivity of 0.89, specificity of 0.95 and overall correct prediction of 92%. The demonstrated predictive capabilities of the cat-SAR approach, together with its modeling flexibility and design transparency, suggest the potential for its widespread applicability to toxicity prediction and for deriving mechanistic insight into toxicologic effects.

Keywords: Structure–activity relationship (SAR); *In silico* modeling; Respiratory sensitizer; Predictive toxicology; Chemical fragments; Categorical SAR (cat-SAR) program

1. Introduction

The task of identifying toxic agents is not a small or trivial challenge. One approach has been to use mathematical models that relate biological activity to chemical structure. Benfenati and Gini [1] describe modern structure–activity relationship (SAR) and quantitative SAR (QSAR) methods as typically involving three parts: (1) the chemical part, (2) the biological part (i.e. activity) and (3) the methodology for relating parts 1 and 2. The main premise for these methods is that recurring and identifiable attributes of chemicals are associated with, or responsible for, particular biological effects. The attributes can take many forms including

*Corresponding author. Email: arc@lsu.edu

SAR and QSAR in Environmental Research
ISSN 1062-936X print/ISSN 1029-046X online
http://www.tandf.co.uk/journals
DOI: 10.1080/10659360500036976

chemical structures, chemicophysical or quantum mechanical properties and graph indices, to name a few. There are numerous methods that relate chemical structure with activity such as those based on human expertise like Ashby's "structural alerts" for potential carcinogenicity [2–4] to statistical QSAR methods like Hansch analysis (see e.g. [5]), comparative molecular field analyses (CoMFA) [6] and MCASE [7–9].

Advances in computing and chemoinformatics, standardized biological or toxicological testing, and the subsequent development of large libraries of test results have ushered in the era of computational or *in silico* SAR. Computational SAR models have gained recent acceptance in the regulatory community for both human health [10] and ecological endpoints [11]. Dearden succinctly summarized the field of computational SAR or *in silico* toxicity prediction to include QSAR models of congeneric and noncongeneric datasets and "expert systems" [12]. The utility and application of some important expert system toxicology prediction methods have been reviewed by Richard [13,14]. Through the use of various techniques, the overall goal is to identify meaningful associations between activity and chemical structure. These associations can then be used to investigate the underlying mechanisms of toxicity, or be extended to estimate or predict the toxicity of untested compounds.

With today's fast CPUs, abundant amounts of computer memory, and the availability of chemical informatics and graphics software we have aimed to readdress the challenge of computer-based SAR expert systems for modeling large and chemically diverse datasets. We describe herein the first generation of a new data and information-intensive approach to toxicological SAR modeling. The program is based on the well-established premise in SAR modeling that like structure begets like activity and employs chemical substructures to differentiate between categories of biologically active and inactive compounds for toxicological endpoints. We have named the new program cat-SAR for categorical SAR.

The cat-SAR program uses 2-dimensional chemical fragments generated by the Sybyl HQSAR module. We chose early in the development process of cat-SAR to use the Sybyl platform which already possessed the needed utilities of *in silico* chemical fragmenting, molecular graphics, and chemical informatics and database requirements associated with our modeling goals. Of importance, the HQSAR module is used solely to generate molecular fragments and is not used for further model development or statistical analysis.

Briefly, the HQSAR module is used to generate a list of chemical fragments associated with compounds in a learning set and produce a data matrix of compounds and fragments. In the data matrix, the rows are the chemicals and the columns are the molecular fragments. Thus for each chemical, a tabulation of all its fragments are recorded across the table rows and for each fragment all chemicals that contain in it are tabulated down the table columns. The compound-fragment matrix is then analyzed, in conjunction with the known biological activity category of each compound, by the cat-SAR program. The cat-SAR program identifies structural features associated with the biologically active and inactive categories. The cat-SAR program, the respiratory sensitizer learning set (described below), and the compound-fragment matrix are available through the corresponding author.

Since cat-SAR modeling is independent of the biological data used in the process we anticipate that it can be generally applied from the study of drugs to environmental toxicants. Moreover, the models can be used for either mechanistic studies of biological phenomena or for the prediction of biological activity for untested compounds.

The cat-SAR program stands alone from other computerized SAR expert systems in its openness, flexibility, routine for identifying important attributes of biological activity or inactivity, and its method for predicting the activity of untested compounds. Several commercially available computational SAR expert systems including MultiCASE, TOPKAT, and Oncologic are relatively closed systems where proprietory (and unknown) routines are used to generate the final model. On the other hand, cat-SAR is completely open with every detail of modeling transparent to the user. As for inflexibility, many of the commercially available expert systems maximally only allow the user to alter the makeup of the learning sets (users cannot alter the parameters for model development). The cat-SAR approach allows the user to select and/or adjust many parameters during the model process from learning set makeup, to selection of types of fragment attributes to consider, to ultimately what numerical or statistical considerations are employed in developing the final model. These are described in detail below.

The cat-SAR approach is also a very data- and information-intensive SAR expert system. During model development and the creation of the final model, all fragments associated with the categories are presented. This leaves the user with an unbiased view of all important features associated with the biological endpoint. Consider the fact that the published MCASE model of the same respiratory sensitizer learning set used herein produced a model based on eight biophores and no biophobes [15]. One of the models developed with the cat-SAR program produced 1213 fragments associated with activity and 92 associated with inactivity. Similarly, the prediction of the activity of compounds outside the model's learning set presents the user with a *complete* correspondence between all the fragments in the model (e.g. 1213 active and 92 inactive) and those in the compound being predicted. Again considering the published MultiCASE report for this dataset, MultiCASE predicted the activity of methyldopa and presented the user with two reasons (i.e. biophores) for why the compound was predicted active. The cat-SAR program provided 22 reasons.

The approach we have taken in developing cat-SAR clearly diverges from existing SAR expert systems and is more in tune with modern QSAR techniques. For instance, the user is presented with a number of selectable and adjustable modeling parameters. The notion of having selectable and adjustable modeling parameters facilitates that ability to rigorously explore the relationships between chemical structure and biological activity.

We chose to test the method on a previously published respiratory sensitization model due to its small size (i.e. 80 compounds) and good modeling potential that was previously demonstrated using CASE-MultiCASE [15]. This model has recently been reviewed by Rodford *et al.* [16].

2. Materials and methods

2.1 Description of the cat-SAR SAR program

The cat-SAR models are built through a comparison of structural features found amongst the active and inactive compounds in the model's learning set. A categorical approach is used with, in this instance, compounds designated as active or inactive. For this exercise, active compounds were chemical respiratory sensitizers and inactive compounds were nonsensitizers. The modeling process began with the compilation of a set of chemicals and their biological activity (described below). Using the Tripos Sybyl HQSAR module, each

chemical was fragmented into all possible fragments. HQSAR allows the user to select attributes for fragment determination including atom size, bond types, atomic connections, inclusion of hydrogen atoms, chirality and hydrogen bond donor and acceptor atoms. Moreover, fragments can be linear, branched or cyclic moieties.

We developed two sets of fragments from the model's learning set. The first (fragment set ABC) contained fragments between three and seven atoms in size and considered Atoms, Bond types, and atomic Connections (i.e. the arrangement of atoms in the fragment). The second (fragment set ABCH) included the same descriptors as the previous set plus associated Hydrogen atoms. A compound-fragment matrix was produced for both sets of fragments.

A measure of each fragment's association with biological activity was next determined. This step is controlled by the user. To ascertain an association between each fragment and activity (or lack of activity) a set of rules is established to choose "important" active and inactive fragments. It should be noted that in this generation of the program we are using a common-sense approach, rather than statistical analysis, to select "significant" fragments.

The first selection rule is the number of times a fragment is identified in the learning set. For this exercise, it was arbitrarily set at three compounds (or 3.75% of the compounds in the learning set). This was a reasonable decision considering that if a fragment is found in only one or two compounds in the learning set it may be a chance occurrence. We do, however, note that fragments found in only one or two compounds may not be outliers but rather underrepresented descriptors of activity. On the other hand, since the learning set is composed of only 40 active and 40 inactive compounds (see next section), if we required fragments to be found in more than three compounds, we would expect to miss important features.

The second rule relates to the proportion of active or inactive compounds that contain each fragment. For both the ABC and ABCH fragment sets, we set the proportion at 0.90. We reasoned that even if a particular fragment is associated with activity, there may yet be other reasons (i.e. fragments) for its being inactive, thus it would not be expected to be found in 100% of the active compounds. Likewise is true for inactive fragments. Thus, if we considered only those fragments found exclusively in active or inactive compounds we would rarify the fragments pool to an unreasonable level and risk losing valuable information. On the other hand, we expected that fragments found to be present approximately equally in the active and inactive fragment sets would not be associated with biological activity. Such fragments may serve as chemical scaffolds holding the biologically active features and are not directly related to activity or inactivity.

In summary, fragments were considered "significant" if they were found in at least three compounds in the learning set and also found in at least 90% of the active or inactive compounds that derived them.

The resulting list of fragments can then be used for mechanistic analysis, or to predict the activity of an unknown compound. In the latter circumstance, the model determines which, if any, fragments from the model's learning set the compound contains. If none are present, no prediction of activity is made for the compound. If one or more fragments are present, the number of active and inactive compounds containing each fragment is determined. The probability of activity or inactivity is then calculated based on the total number of active and inactive compounds containing the fragments.

The probability of activity of a test chemical is calculated from the average probability of active and inactive fragments. For example, if a test compound contains two fragments, one present in 9/10 active compound (i.e. 90% active) and one in 3/3 inactive one

(i.e. 100% inactive), the unknown compound will be predicted to be *active* based on the higher probability of activity derived from chemicals containing these fragments.

In this manner, the probability of activity or inactivity is determined by comparison of the structure of the unknown compound with the entire structural information present in the model.

It requires noting that cat-SAR predictions are based on what can be conceived as two separable models: The inactive fragment model and the active fragment model. By so doing, cat-SAR predictions are based on information that is associated with biological activity and inactivity. The cat-SAR program does not employ the use of default predictions wherein, as in the case of MultiCASE, if no biophores are present in an unknown chemical it is predicted by default to be inactive. This, of course, presents the situation wherein the cat-SAR program will not make predictions on some chemicals. Although this may seem like a drawback to the program by appearing less universal, the user of the program always has the option to simply define chemicals that are not predictable by cat-SAR with a default value.

2.2 Respiratory sensitization databases

The dataset of respiratory sensitizers has been reported by Graham *et al.* [15]. Briefly, chemical sensitizers were identified through a search of the medical literature. Selection criteria were in accordance with the US Department of Health and Human Services "Guidelines for Diagnosis and Treatment of Asthma" [17]. The search criteria included chemicals with inhalation challenge followed by a drop of $>20\%$ in forced expiration volume at 1 s within 24 h of challenge. Forty compounds were identified. No reports were identified of chemicals tested as described and found to be nonsensitizers in humans except for the often-used control substance, lactose. Since, as discussed, the cat-SAR method requires a comparison of biologically active with inactive compounds, we designated as "negative" a set of 40 chemicals previously selected as respiratory nonsensitizers by Graham *et al.* [15]. These 40 compounds were randomly selected from a dataset of chemicals tested for human allergic contact sensitizing ability via patch testing and were found to be nonsensitizers [18]. The assumption was made that dermal nonsensitizers would also be respiratory nonsensitizers. In general, chemicals were relatively small organic compounds that did not include salts, metals, mixtures, or polymers.

3. Results and discussion

3.1 Predictive performance of the cat-SAR respiratory sensitization models

To evaluate the predictive ability of the models, a leave-one-out cross-validation test was conducted. For each chemical in the learning set, one at a time, its chemical fragments were removed from the total fragment set, and the probability of activity or inactivity associated with each fragment was recalculated. Using the criteria described above to estimate activity of unknown compounds, the activity of the removed chemical was predicted.

Overall, the ABC and ABCH models correctly classified 91 and 92% of the chemicals they were capable of predicting (table 1). The predicted activity for each chemical is listed in table 2. The cat-SAR program, using the n-1 cross-validation learning sets (i.e. models built on 79 compounds), was unable to make predictions for five chemicals in the ABC model and

Table 1. Predictive performance of ABC and ABCH respiratory sensitization models. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments**	*Model Fragments†*	*Active Fragments‡*	*Inactive Fragments¶*	*Sensitivity§*	*Specificity‖*	*OCP#*
ABC	5737	1305	1213	92	0.94	0.87	0.91
ABCH	14424	3356	2926	430	0.89	0.95	0.92

*number of fragments derived from learning set.
†number of fragments meeting specified rules of the model.
‡number of fragments meeting specified rules to be considered as active.
¶number of fragments meeting specified rules to be considered as inactive.
§number of correct positive predictions / total number of positives.
‖number of correct negative predictions / total number of negatives.
#Observed Correct Predictions: Number of correct predictions / total number of predictions.

three in the ABCH (table 2). The reason for this is that each of these compounds did not possess any structural features that the n-1 models could base a prediction upon. A previous CASE/MultiCASE model of the same data reported an overall correct classification of 95%. This was based on the Bayesian combination of four CASE/MultiCASE submodels that individually had sensitivities ranging from 72–80% and specificities ranging from 95–98% [15]. In a separate published model based on chemicophysical parameters, a sensitivity of 85% and a specificity of 74% was achieved [19]. Interestingly, the individual ABC and ABCH cat-SAR models are quite balanced with respect to sensitivity and specificity (table 1). This is not the case with the previous CASE/MultiCASE and chemicophysical models. The individual CASE/MultiCASE models tended to have a better ability to predict the inactive chemicals and the chemicophysical model was better able to predict the active ones.

The question arises as to why the program produced wrong predictions. In the case of any of the previously mentioned respiratory sensitizing models, the simplest explanation lies in the possibility that some of the information on which the models were built is not correct. Consider the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed for reproducibility and accuracy by replicate analyses of chemicals [20]. The interlaboratory reproducibility of the *Salmonella* mutagenicity assay is only 85% [20]. Therefore, the databases may contain some incorrect information.

However, other explanations should be considered. The incorrect ABC model prediction for hexamethylene diisocyanate and the incorrect ABC and ABCH model predictions for isophorone diisocyanate are of interest. They both contain the isocyanate moiety which is clearly associated with biological activity. The cat-SAR program also identifies this moiety in these two compounds. However, the compounds contain a number of inactivating fragments that counterbalance the isocyanate-related ones. At this time, a complete understanding of the inaccurate predictions is not possible, but further development of both the models and the databases should lead to a more comprehensive analysis.

3.2 Respiratory sensitization model analysis

As described above, two models were developed using the same set of 80 compounds. These models can be considered as independent since they are built upon different fragment bases. The ABC model started with a total fragment set of 5737 and the ABCH model with a set of

Table 2. Model validation for respiratory sensitizers. Compounds with values above 50% were predicted to be active compounds and those below 50% were predicted to be inactive.

		Model 3-7/3/0.90	
Chemical	*Experimental Activity*	*ABC % Active*	*ABCH % Active*
1,5-Napthalene diisocyanate	+	1.00	1.00
2-(*N*-Benzyl-*N*-*tert*-butylamino)-4′-hydroxy-3′-hydroxymethyl acetophenone diacetate	+	0.63	0.59
2,4-Toluene diisocyanate	+	1.00	1.00
2,6-Toluene diisocyanate	+	1.00	1.00
6-Amino penicillanic acid	+	1.00	1.00
7-Amino cephalosporanic acid	+	0.99	0.99
Ampicillin	+	1.00	1.00
Azocarbonamide	+	1.00	0.98
Benzylpenicillin	+	1.00	1.00
Brilliant orange GR	+	1.00	1.00
Carminic acid	+	0.57	0.54
Cephalexin	+	1.00	1.00
Chlorhexidine	+	1.00	0.96
Dichlorvos	+	*	*
Dimethyl ethanolamine	+	1.00	1.00
Diphenyl methane-4,4′-diisocyanate	+	1.00	1.00
Epigallocatechin gallate	+	0.57	0.60
Ethanolamine	+	1.00	1.00
Ethyl cyanoacrylate	+	*	0.03†
Ethylenediamine	+	1.00	1.00
Fenthion	+	0.91	0.96
Hexamethylene diisocyanate	+	1.00	0.38†
Isononanoyl oxybenzene sulfonate	+	0.98	0.82
Isophorone diisocyanate	+	0.22†	0.17†
Maleic anhydride	+	1.00	1.00
Methyl-2-cyanoacrylate	+	*	*
Methyldopa	+	0.99	0.95
Phenylglycine acid chloride	+	1.00	1.00
Phthalic anhydride	+	1.00	1.00
Piperacillin	+	1.00	1.00
Piperazine	+	1.00	1.00
Plicatic acid	+	0.53	0.74
Reactive orange 3R	+	1.00	1.00
Rifafix red BBN	+	1.00	1.00
Rifazol black GR	+	1.00	1.00
Tetrachloroisophthalonitrile	+	*	*
Tetrachlorophthalic anhydride	+	1.00	1.00
Triethylenetetramine	+	1.00	1.00
Trimellitic anhydride	+	1.00	1.00
Tylosin	+	0.14†	0.14†
1,1,3,3,5-Pentamethyl-4,6-Dinitroindane	−	0.00	0.00
1,4-Cineole	−	0.00	0.04
1-Hexanol	−	*	0.07
2,4-Dimethylbenzyl acetate	−	0.00	0.02
2-Butyl-4,4,6-trimethyl-1,3-dioxane	−	1.00†	0.50
2-*tert*-Amylcyclohexyl acetate	−	0.03	0.06
3,6-Dimethyloctan-3-yl acetate	−	0.05	0.06
3-Butyl phthalide	−	0.03	0.06
4-Acetyl-6-*tert*-butyl-1,1-dimethylindane	−	0.00	0.06
5-Methyl α-ionone	−	0.12	0.09
9-Decenyl acetate	−	0.05	0.05
Acetyl ethyltetramethyltetralin	−	0.00	0.00
Allyl heptylate	−	0.10	0.05
Benzyl butyrate	−	0.10	0.06
Butyl isobutyrate	−	0.06	0.07
Camphene	−	0.00	0.04
cis-3-Hexenyl anthranilate	−	0.65†	0.35

Table 2 – *continued*

Chemical	Experimental Activity	Model 3-7/3/0.90 ABC % Active	Model 3-7/3/0.90 ABCH % Active
cis-4-Decen-1-al	–	0.03	0.04
Citronellyl nitrile	–	0.03	0.05
Cyclohexylethyl alcohol	–	0.00	0.06
Dibutyl sulphide	–	1.00†	0.93
Dihydro-isojasmone	–	0.03	0.04
Dimethylheptenol	–	0.03	0.05
Ethyl acetoacetate ethylene glycol ketal	–	0.27	0.19
Ethyl lactate	–	0.09	0.07
Eugenyl phenylacetate	–	1.00†	0.81†
γ-Dodecalactone	–	0.05	0.07
Geranyl benzoate	–	0.03	0.06
Heptyl butyrate	–	0.06	0.06
Hexane	–	0.00	0.09
Hexyl tiglate	–	0.04	0.06
Isoamyl butyrate	–	0.06	0.06
Lactoscatone	–	0.04	0.05
l-Carvyl propionate	–	0.04	0.04
Methyl tiglate	–	0.09	0.07
Musk xylol	–	0.00	0.00
Phenylethyl acetate	–	0.77†	0.32
p-Isopropylcyclohexanol	–	0.00	0.04
Rhodinyl formate	–	0.03	0.05
Undecenyl acetate	–	0.05	0.05

* no prediction was made for the compound.
† wrong prediction was made for the compound.

14424 fragments (table 1). In both models, approximately 23% of the total number of fragments met the criteria to be considered "significant" (i.e. 1307 significant /5753 total = 22.7% for ABC and 3356 significant /144424 total = 23.2%) (table 1). The remaining fragments were either not present in a sufficient number of compounds (i.e. found in <3 or 3.75% of compounds in the learning set), or the fragments did not come from compounds that were predominately (i.e. $>90\%$) active or inactive.

Overall, both models performed similarly. However, when considering the sensitivity and specificity of the models, the distinction was not clear-cut. The ABC model was better able to correctly predict the active chemicals while the ABCH model was better able to predict the inactive ones. At this point, we chose to focus on the ABC model. This decision was based on several criteria: (1) Both models have nearly equivalent correct prediction rates (table 1) and make similar predictions on the majority of compounds in the validation set (table 2), (2) Considering the law of parsimony, the ABC model is based on fewer fragments and (3) The models are constructed from a set of 40 chemicals *tested* and found to be respiratory sensitizers, whereas the set of 40 chemicals designated as "inactive" are *presumed* to lack activity. Therefore, based on the quality of information of these active and inactive sets, we favored a model with better ability to predict activity as compared with inactivity.

Although beyond the scope of this report, we bring attention to the finding that the cat-SAR method derives multiple independent models for the same endpoint. The observation that the ABC and ABCH models do not predict the same activity for each chemical suggests that the models may be capable of describing different attributes of the activity. This suggests

the possibility of development of a consensus model using a Bayesian technique similar to those previously reported using CASE/MultiCASE [15].

3.3 Examples of the cat-SAR model predictions

Methyldopa and 2,4-dimethylbenzyl acetate were selected to demonstrate the predictive ability of the cat-SAR modeling method for an active and inactive chemical, respectively. For this demonstration, we used the ABC model for reasons just described. Tables 3 and 4 list the significant fragments derived from the two compounds. Figures 1 and 2 illustrate the intact compounds and their associated fragments. The predictions presented for the two compounds are based on results obtained from the leave-one-out validation exercise. Therefore, the compounds themselves are not contributing to the fragment set of the model and are thus not influencing their own prediction of activity or inactivity.

Table 3 lists and figure 1 shows all the significant fragments used in the leave-one-out validation exercise to predict the activity of methyldopa. Methyldopa was predicted to have a probability of activity of 0.988. This represents the average probability of activity of the 22 fragments used in the prediction (table 2). No fragments associated with methyldopa were considered inactive.

Likewise, table 4 and figure 2 shows all the significant fragments used in the validation exercise to predict the activity of 2,4-dimethylbenzyl acetate. 2,4-Dimethylbenzyl acetate was predicted to have a probability of inactivity of 1.0.

As indicated, the prediction for the respiratory sensitizing ability of methyldopa and 2,4-diemthylbenzyl acetate were based on the complete correspondence of significant fragments

Table 3. Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory sensitizer methyldopa

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
frag258	10	1	11	0.909	0.091
frag283	10	1	11	0.909	0.091
frag308	10	1	11	0.909	0.091
frag348	8	0	8	1.000	0.000
frag357	8	0	8	1.000	0.000
frag400	14	0	14	1.000	0.000
frag471	6	0	6	1.000	0.000
frag522	6	0	6	1.000	0.000
frag914	4	0	4	1.000	0.000
frag915	4	0	4	1.000	0.000
frag920	4	0	4	1.000	0.000
frag921	4	0	4	1.000	0.000
frag2378	3	0	3	1.000	0.000
frag2401	3	0	3	1.000	0.000
frag2415	3	0	3	1.000	0.000
frag2416	3	0	3	1.000	0.000
frag2463	3	0	3	1.000	0.000
frag2471	3	0	3	1.000	0.000
frag2472	3	0	3	1.000	0.000
frag2507	3	0	3	1.000	0.000
frag2509	3	0	3	1.000	0.000
frag2706	3	0	3	1.000	0.000
	Probability of activity			0.988	0.012

* number of active compounds that contain the fragment.
† number if inactive compounds that contain the fragment.
‡ number of compounds in the dataset that contain the fragment.

Table 4. Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory nonsensitizers 2,4-Dimethylbenzyl acetate.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
frag4970	0	3	3	0.000	1.000
frag4979	0	3	3	0.000	1.000
frag4982	0	3	3	0.000	1.000
frag5003	0	4	4	0.000	1.000
frag5011	0	4	4	0.000	1.000
frag5032	0	4	4	0.000	1.000
frag5033	0	4	4	0.000	1.000
frag5073	0	4	4	0.000	1.000
	Probability of activity			0.000	1.000

See table 3 footnotes for reference.

from the model's validation set to all the fragments identified in the compound. Methyldopa was predicted to be active based on 22 fragments from its validation set of fragments. Inspection of these fragments revealed several major themes. Fragment 348 leads to a series of complimentary moieties covering the amine to carboxylic acid portion of the molecule. Fragment 283 covers the *para* unsubstituted phenol and accounts for four other validation fragments. Fragment 2706

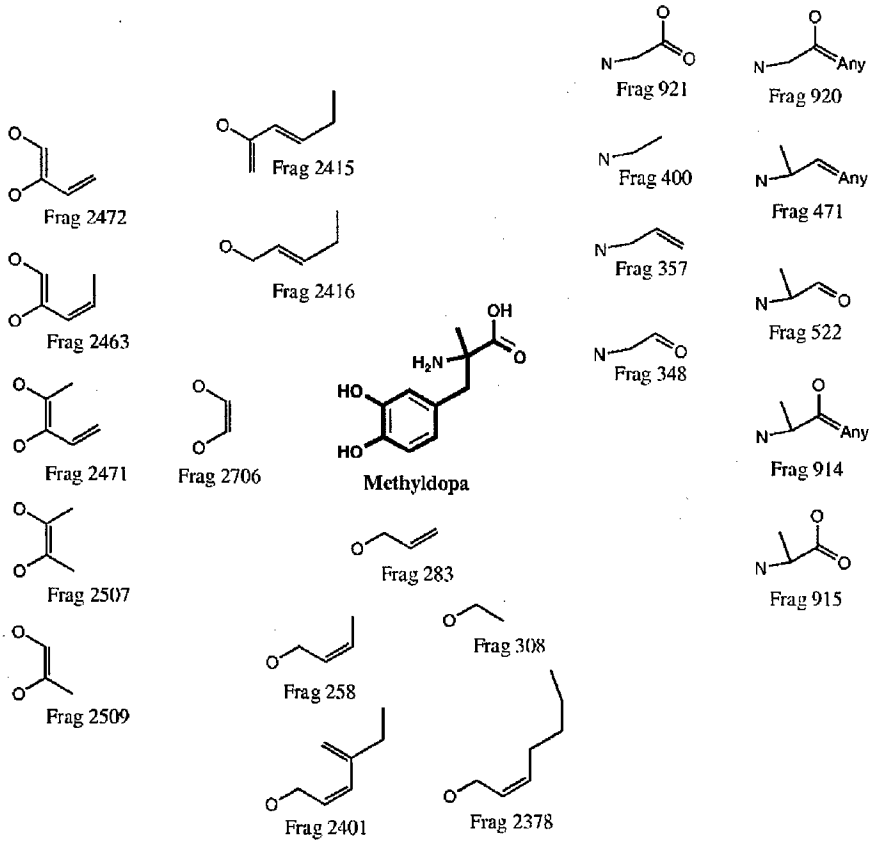


Figure 1. Illustration of the 22 significant fragments contributing to the active validation prediction of methyldopa.

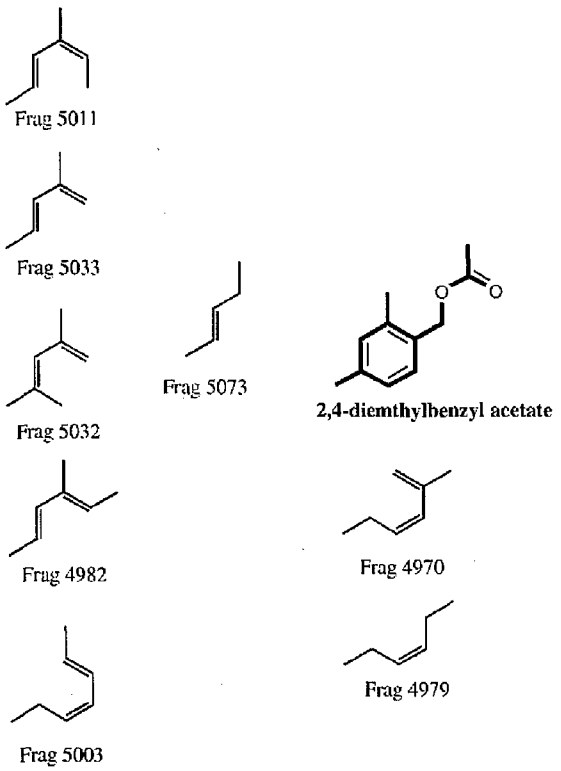


Figure 2. Illustration of the eight significant fragments contributing to the inactive validation prediction of 2,4-dimethylbenzyl acetate.

covers the 3,4-diol and accounts for five other validation fragments. Fragments 2415 and 2416 are closely related to Fragment 2706 but cover just the 3-hydroxyl.

For 2,4-dimethylbenzyl acetate, Fragments 4970 and 4979 cover the *para* substituted methyl section of the molecule. Moreover, Fragment 5073 covers the 2,4-methyl substitution and can account for four similar fragments.

From a prediction point-of-view, any one fragment would have been sufficient for the accurate prediction in these examples. From a mechanism point-of-view, for methyldopa, just the four major fragment families (i.e. from fragments 348, 283, 2706, and 2416) would have covered the major identified structural themes relating to activity. The same is true for 2,4-dimethylbenzyl acetate where two sets of similar fragments (i.e. from fragments 5073–4970) described the compound. In this model, the fragment redundancy is obvious. However, we speculate that this may not be the case with other toxicological endpoints. In models for other endpoints, where fragments are similar but not exact, each fragment may contribute novel mechanistic and predictive information to the model.

Clearly, from the results of the validation exercises, the cat-SAR program in not performing at 100% accuracy. To judge the predictive performance of our models, we compared them to two previously developed MCASE models. One model is based on the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed

for reproducibility and accuracy by replicate analyses of chemicals [20]. As previously indicated, the interlaboratory reproducibility of the *Salmonella* mutagenicity assay is only 85% [20].

4. Conclusions

The new cat-SAR modeling approach described herein has a predictive ability in line with other respiratory sensitization models developed by us [15,19]. This clearly suggests its utility and warrants further development. It is applicable to toxicological or pharmacological SAR modeling. The cat-SAR program uses a binary approach to identify structural features associated with biological activity or inactivity. This is straightforward when the toxicologic endpoint is categorical (e.g. sensitizers vs. nonsensitizers, carcinogens vs. noncarcinogens or mutagens vs. nonmutagens). However, for other endpoints, where a continuous scale of activity is measured, the dichotomy can be imposed between highly active and less active compounds (e.g. extremely toxic vs. nontoxic as in the case of LD_{50} values or high or low receptor affinity as in the case of estrogen receptor ligands).

The cat-SAR method has two main areas of strength when compared with other 2-dimensional modeling systems. The first is the transparency of the method. The derivation of model fragments and decision rules are open for inspection. The entire compound-fragment matrix and the identified model fragments are all easily inspected. The second strength is the amount of user-selectable parameters available for adjustment. For the fragment development part of the program, the user can select fragments of different size and choose other fragment attributes including the consideration of atoms, bond, and hydrogen atoms. Moreover, when identifying important or significant fragments the user can manipulate the selection process by altering the requirements for how many compounds in the learning set contain each fragment and also what proportion of active or inactive compounds in the learning set contain the fragment.

Thus, the cat-SAR method is transparent with regard to the overall modeling process. Users of the program have the opportunity to optimize the process for their own needs. Considering the fact that toxicologic endpoints differ in their mechanisms, it makes sense that the modeling algorithm should be transparent to meet the requirements of the endpoint being modeled.

Overall, in prediction mode, this method presents the user with a *complete* correspondence of fragments in the model and the unknown chemical. In model analysis mode, the method provides the user with a complete listing of all interesting fragments. It should be noted that there is no hierarchy of fragments or filtering of "significant" fragments other than what the user chooses. There are no hidden or proprietary rules in the process. All fragments that meet the user-specified structural requirements and the rules of association with activity or inactivity are included in the model. This leads to the identification of many (e.g. 1000 s) fragments, some with great structural similarity. This clearly presents difficulty in being able to succinctly describe the model. However, important information is retained and accessible to the user.

The cat-SAR program of course has some drawbacks and limitations. Like so many other expert systems in toxicology, it is applicable only to organic chemicals. Metals, mixtures, and polymeric compounds are not suitable for analysis. Moreover, as mentioned, the cat-SAR program presents the final SAR model, in terms of all relevant fragments. This lead to a model that may contain 1000 s of fragments which may lead to difficulty in model interpretation.

Acknowledgements

We gratefully acknowledge support for the development of the cat-SAR program from the Department of Defense Congressionally Directed Medical Research Program for Breast Cancer Idea Award DAMD17-01-0376.

References

[1] E. Benfenati, G. Gini. *Toxicology*, **119**, 213 (1997).
[2] J. Ashby, D. Paton. *Mutat. Res.*, **286**, 3 (1993).
[3] J. Ashby. *Environ. Mutagen.*, **7**, 919 (1985).
[4] J. Ashby, R.W. Tennant. *Mutat. Res.*, **257**, 229 (1991).
[5] C. Hansch, A. Leo. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, D.C. (1995).
[6] R.D. Cramer, D.E. Patterson, J.D. Bunce. *J. Am. Chem. Soc.*, **110**, 5959 (1988).
[7] G. Klopman. *J. Am. Chem. Soc.*, **106**, 7315 (1984).
[8] G. Klopman. *Quant. Struct. Act. Relat.*, **11**, 176 (1992).
[9] G. Klopman, H.S. Rosenkranz. *Mutat. Res.*, **305**, 33 (1994).
[10] M.T.D. Cronin, J.S. Jaworska, J.D. Walker, M.H.I. Comber, C.D. Watts, A.P. Worth. *Environ. Health Perspect.*, **111**, 1376 (2003).
[11] M.T.D. Cronin, J.D. Walker, J.S. Jaworska, M.H.I. Comber, C.D. Watts, A.P. Worth. *Environ. Health Perspect.*, **111**, 1376 (2003).
[12] J.C. Dearden. *J. Comput. Aided Mol. Des.*, **17**, 119 (2003).
[13] A.M. Richard. *Toxicol. Lett.*, **102–103**, 611 (1998).
[14] A.M. Richard. *Knowl. Eng. Rev.*, **14**, 307 (1999).
[15] C. Graham, H.S. Rosenkranz, M.H. Karol. *Regul. Toxicol. Pharmacol.*, **26**, 296 (1997).
[16] R. Rodford, G. Patlewicz, J.D. Walker, M.P. Payne. *Environ. Toxicol. Chem.*, **22**, 1855 (2003).
[17] USDHHS. *U.S. Department of Health and Human Services, National Institutes of Health, Publication No. 90–3042* (1991).
[18] C. Graham, R. Gealy, O.T. Macina, M.H. Karol, H.S. Rosenkranz. *Quant. Struct. Act. Relat.*, **15**, 224 (1996).
[19] M.H. Karol, O.T. Macina, A.R. Cunningham. *Ann. Allergy. Asthma. Immunol.*, **87**, 28 (2001).
[20] W.W. Piegorsrch, E. Zeiger. In *Statistical Methods in Toxicology*, L. Hotham (Ed.), pp. 35, Springer-Verlag, Heidlberg (1991).

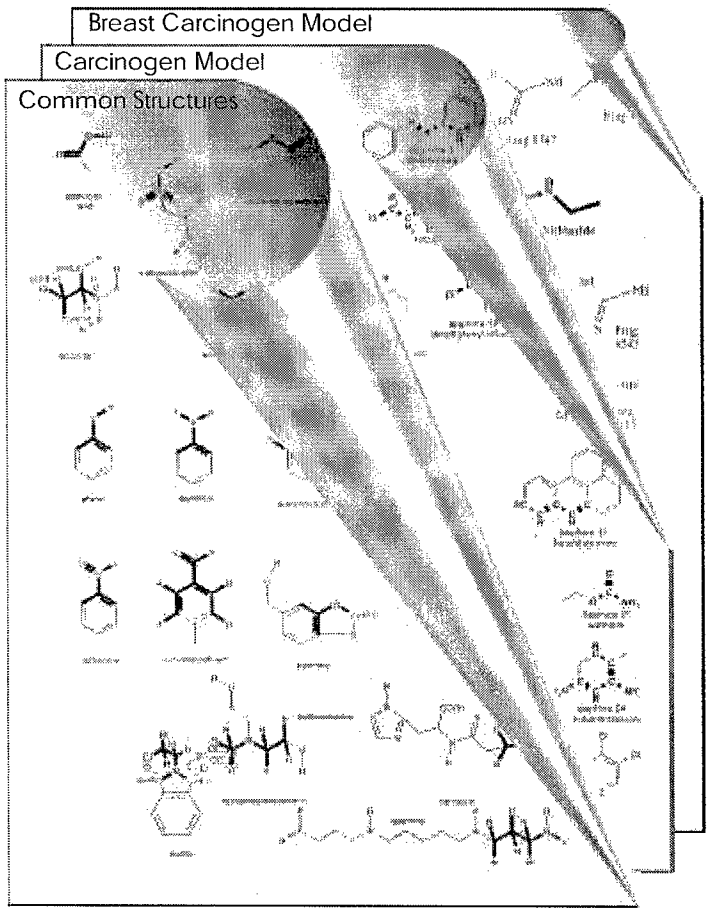
A predictive and mechanistically insightful structure-activity relationship analysis of rat mammary carcinogens

Albert R. Cunningham*, Shanna T. Moss, and Suzanne L. Cunningham

Department of Environmental Studies
Louisiana State University
Baton Rouge, LA 70803

*Corresponding author
1285 Energy, Coast & Environment Building
Baton Rouge, LA 70803
225-578-9422
arc@lsu.edu

Running title: **structure-activity analysis of rat mammary carcinogens**



TOC Graphic Illustration of how SAR models can remove layers of chemical information in order to study specific aspects of the process. Consider chemical carcinogenesis: A SAR model developed from many chemicals, that have been categorized as carcinogens and noncarcinogens removes common chemical structures (top layer) to reveal features associated with carcinogens (middle layer). The SAR model described herein was subsequently developed from carcinogens that have been categorized as breast carcinogens and non-breast carcinogens. This later model removes carcinogen-related structures (middle layer) to reveal a set of features associated with breast-specific carcinogens (bottom layer).

Abstract

Structure-activity relationship (SAR) models are powerful tools to investigate the mechanisms of action of chemical carcinogens and to predict the potential carcinogenicity of untested compounds. We describe herein the application of the recently developed cat-SAR algorithm to two learning sets of rat mammary carcinogens. One set of models developed was based on a comparison of rat mammary carcinogens to noncarcinogens (MC-NC) and the second compared mammary carcinogens to non-mammary carcinogens (MC-NMC). The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88%. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74%. The MC-NMC model was based on a learning set that contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories and was able to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Based on a structural comparison between this model and one for *Salmonella* mutagens, there was no observed relationship between the two phenomena since both the active and inactive categories contained a high proportion of *Salmonella* mutagens. Overall, these findings suggest that the MC-NMC model is identifying structural attributes to address the specific question of "why do some carcinogens cause cancer in the breast?" which is a significantly different question than "why do some chemicals cause cancer?'.

Introduction

The identification of human carcinogens is a difficult and complex task. Only a limited number of high-quality epidemiological studies have been conducted that identify particular agents that induce cancer in humans. In lieu of such data, rodent cancer bioassays or short-term tests for genotoxicity have been used to estimate the likelihood that particular chemicals will be human carcinogens.

However, it is evident that not all chemicals in use today will be tested for carcinogenesis. There are approximately 75,000 industrial chemicals on the Toxic Substance Control Act's Chemical Substance Inventory *(1)* and the National Institute of Environmental Health Sciences estimates that there are over 80,000 chemicals registered for use in the United States *(2)*. A complete two-year bioassay as conducted by the National Toxicology Program (NTP) including planning, evaluation, and review takes about five years to complete, costs between $2-4 million, and uses 400 animals *(3)*. To test all chemicals in this manner is thus prohibitive.

In fact, the NTP has only tested over 500 chemicals for rodent carcinogenicity in standardized 2-year rodent bioassays. Furthermore, the Carcinogenic Potency Database (CPDB) analyzes and consolidates into a single resource the world's diverse literature and NTP Technical Reports of chronic long-term animal cancer bioassays *(4)*. To date, analyses of 6073 experiments on only 1458 chemicals are available on the CPDB's web site *(5)*. Fortunately, the consolidation, standardization, and analyses of cancer bioassay data by the CPDB provides a comprehensive resource for investigating chemical carcinogenesis including analyses by structure-activity relationship (SAR) modeling and predictive toxicological methods.

SAR modeling and other predictive toxicological methods provide a means to estimate toxicological properties of chemicals based on information from previously tested compounds. We have reported predictive and mechanistically insightful SAR models for mice *(6)* and rats *(7)* using the CASE/MULTICASE SAR expert system and chemical carcinogenicity data from the first five plots of the CPDB *(8-12)*. Depending upon validation methods, these models had an observed correct prediction (OCP) rate for chemicals removed from the model's learning set of between 64% and 78% *(6,7)*. Many others have also demonstrated varying degrees of success modeling chemical carcinogens. The utility and application of some important toxicologically-focused predictive methods have been reviewed in-depth by Richard *(13,14)*.

The SAR models of rat and mouse carcinogens developed by us, while being predictive, also provided insight into the structural underpinnings for species-specific carcinogenesis. Many, though not all, of the readily explainable attributes of these models corresponded with the genotoxic or electrophilic paradigm of carcinogenesis *(15)*. In retrospect, this is not surprising given the large numbers of electrophilic or proelectrophilic carcinogens used to build the models and the *a priori* acceptance of the electrophilic theory. These findings did however provide solid evidence that many of the features developed for the models were justifiable and mechanistically sound.

Of note, even in light of this bias toward electrophilicity, we were able to glean an interesting relationship between estrogenicity and carcinogenicity. We identified a 2-dimensional feature of rodent carcinogens that dichotomizes the so-called "beneficial" (e.g., phytoestrogens) from

"harmful" (e.g., pesticides and industrial chemicals) xenoestrogens *(16)*. Further investigation of this feature showed that differences in regional lipophilicity were evident between phytoestrogens and other man-made xenoestrogens. We speculated at the time that these differences in chemical features of estrogen could induce different biological responses *(16-18)*. During this same time, the estrogen receptor alpha (ERα) ligand binding domain was crystallized and its atomic coordinates resolved with those of bound estradiol and raloxifene *(19)*, genistein *(20)*, and 4-hydroxytamoxifen and diethylstilbestrol *(21)*. It was noted that the lipophilic cavity is nearly twice the size of estradiol, which may explain in part the ER's promiscuity *(22)*. Most importantly, it was observed by these authors that estrogen antagonists induce a different conformational change in the AF-2 region compared to that for the natural ligand. These analyses demonstrated the utility of SAR analysis to not only generate predictive models that are explainable by current knowledge but also their ability to provide hypothetical (and testable) information regarding the mechanistic action of toxicants.

The study described herein uses a new SAR algorithm to analyze chemicals that specifically induce mammary cancer in rats. Environmental risk factors, including chemical exposure, may play a role in the development of breast cancer. Unfortunately, many of these factors remain largely unknown. We note that one group of chemicals that has received considerable attention with regards to breast cancer is the environmental endocrine disruptors, with specific attention paid to the xenoestrogens *(23,24)*. For instance, many industrial chemicals (e.g., PCBs and pesticides), consumer products (e.g., plasticizers and phenols), and plant products (e.g., phytoestrogens such as genistein and coumestrol) have been shown to possess estrogenic activity in a number of *in vitro* and *in vivo* assays. As such, xenoestrogens warrant vigorous attention,

especially in light of conflicting epidemiological data and expert opinions regarding their role in the development of this disease *(25-28)*. Overall, in a thorough review of the literature, it was concluded that the available data do not support *or* reject the relationship between exposure to organochlorine compounds and breast cancer *(25)*. As for the role of xenoestrogens in general, the National Research Council states that in fact most studies have been limited primarily to DDT, DDE, TCDD, and PCBs with other compounds receiving little or no attention *(29)*.

Although xenoestrogens have received much public attention, there is a sizable majority of rodent mammary carcinogens that are not estrogenic. Thus, although environmental estrogens may play an important role in the development of breast cancer, other nonestrogenic chemicals may also contribute to the disease. As such, any screening approach designed to identify environmental estrogens, although useful, will not allow for the identification of all potential mammary carcinogens.

The Food and Drug Administration's (FDA) National Center for Toxicological Research recently noted that FDA reviewers are interested in organ-specific carcinogenicity to aid in evaluating new chemicals *(30)*. As such, they have undertaken the task of building an organ-specific database of chemical carcinogens from CPDB data. In their preliminary SAR analyses of liver carcinogens, they obtained a correct prediction rate of 63%, with a sensitivity of 30% and a specificity of 77% *(30)*. Their efforts in developing this comprehensive dataset of organ-specific toxicological data will provide a needed resource for the development and validation of organ-specific SAR models.

Aside from the practical needs of entities like the FDA for estimating organ-specific toxicity, the development of organ-specific carcinogenicity models is also technically appealing. SAR models developed from whole-animal carcinogenicity data attempt to deal with many underlying and often competing mechanisms. As such, it is quite possible that information is lost in the modeling process. Therefore by focusing on specific organs, we hypothesize the development of a clearer picture of the chemical requirements for carcinogenesis in that organ.

For the analyses described herein, we used a newly developed SAR expert system to analyze the set of rat mammary carcinogens reported in the CPDB *(31)*. The system is called cat-SAR for categorical-SAR. Basically, the cat-SAR approach is a computational SAR or *in silico* toxicity prediction "expert system" as classified by Dearden *(32)*. In a previous analysis of human respiratory sensitizers, the cat-SAR program was able to achieve an overall correct prediction rate of 92% with sensitivities between 89 and 94% and specificities between 87 and 95% *(33)*.

The approach we have taken in developing the cat-SAR program clearly diverges from existing commercial SAR expert systems and is more in tune with modern QSAR techniques. For instance, the user is presented with a number of selectable and adjustable modeling parameters. The control and selection of modeling parameters facilitates the ability to rigorously explore the relationships between chemical structure and biological activity. Ultimately, this rationale negates any *a priori* requirements that a given set of data must fit the attributes of a predefined and often proprietary modeling process.

The cat-SAR models are built through a comparison of structural features found amongst categorized compounds in the model's learning set. Generically, these categories are biologically active and inactive compounds. When just considering whole animal carcinogenesis, the categories are simply carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis, the question arises as to the selection of the inactive or noncarcinogenic compounds. Should they be whole animal noncarcinogens or carcinogens that are just not carcinogenic to the organ under consideration? For this exercise, we considered both options and developed predictive SAR models comparing rat mammary carcinogens to noncarcinogens (MC-NC model) and rat mammary carcinogens to non-mammary carcinogens (MC-NMC model).

Materials and Methods

Mammary Gland Carcinogen Learning Sets

The CPDB standardizes the experimental results (whether positive or negative for carcinogenicity), including qualitative data on strain, sex, route of compound administration, target organ, histopathology, and the author's opinion and reference to the published paper, as well as quantitative data on carcinogenic potency, statistical significance, tumor incidence, dose-response curve shape, length of experiment, duration of dosing, and dose rate *(8)*. Moreover, a potency value for carcinogens, the TD_{50} is available. The TD_{50} is "that dose rate (in mg/kg body weight/day) which, if administered chronically for the standard lifespan of the species, will halve the probability of remaining tumorless throughout that period" *(8)*.

The rat mammary carcinogen learning sets were developed from the published CPDB carcinogen target site summary *(31)*. This reference listed 102 rat mammary carcinogens. We excluded norlestrin and dimethylaminoethylnitrosoethyl urea nitrite salt. Norlestrin is a mixture, while the second compound is an organic complex. Therefore a total of 100 rat mammary carcinogens were included in the learning sets.

As discussed below, the cat-SAR program derives SAR models through the comparison of structural features associated with categorical responses (e.g., active *vs.* inactive compounds or carcinogens *vs.* noncarcinogens). When just considering whole animal carcinogenesis, the categories are simply carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis, selection of the inactive or noncarcinogenic compounds could be whole animal noncarcinogens or carcinogens that are not carcinogenic to the organ under consideration. We considered both options for this analysis. Hence, we developed two separate sets of models for rat mammary carcinogens: The mammary carcinogen - noncarcinogen (MC-NC) model and the mammary carcinogen - non-mammary carcinogen model (MC-NMC).

Since we had sufficient numbers of noncarcinogens and non-mammary carcinogens to include as "inactives", we made triplicate inactive datasets (designated Sets 1, 2 and 3 in Table 1) of 100 chemicals each. By so doing we were able to assess the stability of the derived models. Statistical comparison of the each of the models' fragment sets and predictivity was conducted to determine whether the three sets were statistically different. Moreover, this approach prevented the chance of selecting 100 inactive compounds that produced a "good" model. For the MC-NC model three random sets of 100 noncarcinogens were randomly selected from the 449 rat

noncarcinogens listed in the CPDB. Likewise, for the MC-NMC model three random sets of 100 carcinogens were selected from the 395 rat carcinogens in the CPDB that did not induce mammary cancer.

The Categorical-SAR or cat-SAR Expert System Methodology

The Learning Set

The cat-SAR models are built through a comparison of structural features found amongst two designated categories of compounds in the model's learning set. As mentioned, for these analyses the categories for the first model were mammary carcinogens *vs.* noncarcinogens (MC-NC) and mammary carcinogens *vs.* non-mammary carcinogens (MC-NMC) for the second model. The cat-SAR learning set consists of the chemical name, its structure as a .MOL file, and its categorical designation (e.g., one or zero). Organic salts are included as the freebase. Simple mixtures and technical grade preparations are included as the major or active component. Metals, metaloorganic compounds, polymers, and mixtures of unknown composition are not included.

Since the cat-SAR program requires a number of user-specified options, there is not an *a priori* determination of the final model. In other words, the user is allowed to explore and optimize the modeling program. As such, we have developed and reported herein several different cat-SAR MC-NC and MC-NMC models.

In Silico Chemical Fragmentation and the Compound-Fragment Data Matrix

Using the Tripos Sybyl HQSAR module, each chemical was fragmented *in silico* into all possible fragments meeting user-specified criteria. HQSAR allows the user to select attributes for fragment determination including atom count (i.e., size of the fragment), bond types, atomic connections (i.e., the arrangement of atoms in the fragment), inclusion of hydrogen atoms, chirality and hydrogen bond donor and acceptor groups. Fragments can be linear, branched or cyclic moieties.

The first sets of models developed contained fragments between three and seven atoms in size and considered Atoms, Bond types, and atomic Connections. These are referred to as ABC fragment set models. The second set included the same descriptors as the previous set plus associated Hydrogen atoms. These are referred to as ABCH fragment set models.

Upon completion of the fragmentation routine, a Sybyl HQSAR add-on procedure produces the compound-fragment data matrix as a text file. In the matrix, the rows are intact chemicals and the columns are molecular fragments. Thus for each chemical, a tabulation of all its fragments are recorded across the table rows and for each fragment all chemicals that contain it are tabulated down each column.

The HQSAR module is not used for statistical analysis or model development. The compound-fragment matrix is then analyzed, using the cat-SAR programs we have developed in order to identify structural features associated with active and inactive compounds. The programs, mammary carcinogen models, and the compound-fragments matrix are available through the corresponding author.

Identifying "Important" Fragments of Activity and Inactivity

A measure of each fragment's association with biological activity was next determined. This step is controlled by the user. To ascertain an association between each fragment and activity (or inactivity) a set of rules is established to choose "important" active and inactive fragments. It should be noted that in this generation of the program we are using a common-sense approach, rather than statistical analysis, to select "significant" fragments.

The first selection rule is the number of times a fragment is identified in the learning set. For this exercise, it was arbitrarily set at three compounds in the learning set (i.e., 1.5%). This was a reasonable decision considering that if a fragment is found in only one or two compounds in the learning set it may be a chance occurrence. We do, however, note that fragments found in only one or two compounds may not be outliers but rather underrepresented descriptors of activity. On the other hand, since the learning sets are composed of 100 active and 100 inactive compounds, if we required fragments to be found in more than three compounds, we would expect to miss important features.

The second rule relates to the proportion of active or inactive compounds that contain each fragment. We derived models for two proportions, 0.75 and 0.90, for both the ABC and ABCH fragment sets. In general, we reasoned that even if a particular fragment is associated with activity, there may be other reasons (i.e., fragments) for its being inactive, thus it would not be expected to be found in 100% of the active compounds. A similar argument can be made for inactive fragments. Thus, if we considered only those fragments found exclusively in active or

inactive compounds, we would rarify the fragments pool to an unreasonable level and risk losing valuable information. On the other hand, we expected that fragments found to be present approximately equally in the active and inactive fragment sets would not be associated with biological activity. Such fragments may serve as chemical scaffolds holding the biologically active features and are not directly related to activity or inactivity.

In summary, fragments were considered "significant" if they were found in at least three compounds in the learning set and depending on model, also found in at least 75% or 90% of the active or inactive compounds that derived them. The models developed are listed in Table 1.

Predicting Activity

The resulting list of fragments can then be used for mechanistic analysis, or to predict the activity of an unknown compound. In the latter circumstance, the model determines which, if any, fragments from the model's learning set the compound contains. If none are present, no prediction of activity is made for the compound. If one or more fragments are present, the number of active and inactive compounds containing each fragment is determined. The probability of activity or inactivity is then calculated based on the total number of active and inactive compounds containing the fragments.

The probability of activity of a predicted chemical is calculated from the average probability of the active and inactive fragments contained in it. For example, if a compound contains two fragments, one being found in 9/10 active compounds in the learning set (i.e., 90% active) and the other being found in 3/3 inactive compounds (i.e., 100% inactive), the unknown compound

will be predicted to be inactive based on the higher probability of inactivity derived from chemicals containing these fragments. In this manner, the probability of activity or inactivity is determined by comparison of the structure of the unknown compound with the entire structural information present in the model.

"Validating" the Model

The cat-SAR program contains a leave-one-out cross-validation routine. For each chemical in the model's learning set, one at a time, its chemical fragments were removed from the total fragment set, and the probability of activity or inactivity associated with each remaining fragment was recalculated. Using the same criteria described above, the activity of the removed chemical was then predicted using the reduced fragment set.

The cat-SAR predictions are based on two separable fragment sets: The inactive fragments and the active fragments. As mentioned, the predicted activity of a chemical is based on the average probability of all the active and inactive compounds contributing to its fragments. As such, the user can "decide" at what predicted probability of activity or inactivity to categorize the test compound as a carcinogen or noncarcinogen.

To address this, we have adapted a routine from our previous MultiCASE work in which we identify a cut-off point that optimally separates the prediction of active and inactive compounds. This is based on the results of the validation exercise. In other words, since the prediction of activity or inactivity is a probability, we allow the validation exercise to guide us in determining, based on a probability of activity, what is ultimately classified as an active or inactive prediction.

Results and Discussion

Overview of Predictive Performance of the Cat-SAR Mammary Carcinogen Models

The best rat MC-NC model achieved an 82% observed correct prediction (OCP) rate with a sensitivity of 77% and a specificity of 88% (ABC 3/90 Model 1, Table 1). This model made predictions on 145 of the 200 chemicals in the learning set. The best rat MC-NMC model achieved a 79% OCP rate with a sensitivity of 83% and a specificity of 74% (ABC 3/90 Model 2, Table 2). This model made predictions on 124 of the 200 chemicals in the learning set.

In order to better judge how well these two models performed, we can consider the "accuracy" or reproducibility of *in vivo* or *in vitro* toxicological tests themselves. In general, surrogate tests and carcinogen bioassays are not reproducible with 100% concordance. For instance, the NTP's *Salmonella* mutagenicity database, which is derived from a standardized protocol, has been estimated to be 85% reproducible *in vitro (34)*. Moreover, it was found that based on "near-replicate" experiments in the CPDB, there was also a degree of non-reproducibility *(4,35)*. For example, 11 out of 54 chemicals tested in similar experiments for their ability to induce cancer in mice were discordant (i.e., 80% reproducible) and 16 out of 104 chemicals tested for cancer in rats were discordant (i.e., 85% reproducible) *(4)*. This does not imply that the CPDB data is flawed, only that there is variability in results. However, based on these findings, and since the majority of results in the CPDB have not been subjected to replication, it does suggest that SAR models built on this data will not, nor should be expected to achieve 100% accuracy. The cat-SAR mammary carcinogen models appear to be in the same neighborhood of predictivity as the bioassays themselves.

Analysis of Random Subsets of Inactives Models

Statistical analysis of each set of three models derived from the random selection of non carcinogenic or nonmammary carcinogenic compounds indicated the models had approximately the same OCP rate. The most variable set of OCP values was for the rat MC-NMC ABCH 3/90 model where it varied from 72-79% (Table 2). All other models showed a closer spread of OCP values. This provides a degree of confidence that the accurate predictions made by the models were not spurious events based on the fortuitous selection of "good" compounds for the learning set. In other words, this provides assurance that the models are based on a sound foundation and are not providing arbitrary predictions or mechanistic assertions.

Comparison of 75% vs. 90% Models

In order to best compare the 75% and 90% models as well as comparisons described below between the ABC and ABCH models, we chose to consider the average values for fragment counts, OCP, sensitivity, and specificity. These were calculated for each set of three models derived from different sets of randomly selected inactive compounds. This makes discussion of the data more straightforward and also provides for a more robust analysis. However, similar observations can be found when comparing individual models as well.

As the criteria for selecting important fragments increased from those found in 75% of active or inactive compounds to those found in 90%, the model's OCP, sensitivity, and specificity increased. For example, when looking at the ABC MC-NC models, as the fragment selection criteria increased from 75 to 90% the average OCP rose from 74 to 81% (Table 1). The

sensitivity and specificity also rose from 71 to 81% and from 78 to 80%, respectively (Table 1). Similar increases can be seen for all other rat MC-NC and MC-NMC models (Tables 1 and 2).

The trend in improved OCP, sensitivity, and specificity indicates that as the requirements for selecting important fragments are tightened (e.g., increasing proportion from 75% to 90%) the accuracy of predictions made from the resultant more stringent model increased. This is not unexpected. There is a cost associated with this increased accuracy, however. The more stringent models do not contain as many important fragments as the less stringent ones. The rat MC-NC ABC 3/75 model was based on an average of 1484 fragments while the 3/90 model contained 1152 (Table 1). Similar trends can be found for the other models.

Comparison of ABC vs. ABCH Models

Upon comparison of the ABC to ABCH models, the ABC models typically performed better. In general, the ABCH models contained about twice as many fragments. For example, the MC-NC ABC models on average were based on 16979 fragments and the set of ABCH models were based on 36947 fragments (Table 1). Similar increases were also observed for the total number of important fragments as well as the pool of active and inactive fragments. The reason for this is that, as expected, the fragments in the ABCH models are more specific by explicitly considering hydrogen atoms.

Interestingly, these more specific ABCH fragments did not enhance the overall predictivity of the models, but lowered it. For example, the rat MC-NC ABC 3/90 had an average OCP of 81% while the ABCH 3/90 had an OCP of 76% (Table 1). This trend is evident in other comparisons

between the ABC and ABCH models. However, in general the ABCH models were able to make predictions on a greater number of compounds. Again, for example, the rat MC-NC ABC 3/90 model made predictions on 139 chemicals while the ABCH 3/90 made predictions on 164 compounds (Table 1). Similar results were seen for other MC-NC as well as MC-NMC models (Tables 1 and 2).

Examples of cat-SAR Predictions

Atrazine and fenaminosulf were selected to illustrate cat-SAR predictions of mammary carcinogens based on the MC-NC model. Nithiazide and 1-phenyl-3,3-dimethyltriazene were selected to illustrate the MC-NMC models. These "prediction" examples are based on results obtained from the leave-one-out validation exercise. Therefore, the compounds themselves are not contributing to the fragment set of the model and are thus not influencing their own prediction of activity or inactivity. Additionally, the following discussion of the compounds with consideration of their tumor sites and *Salmonella* mutagenicity are based on their classification in the CPDB.

Atrazine is a male mammary carcinogen in rats and does not induce tumors in any other sites in the rat male. Atrazine also induces cancer of the hematopoietic system and uterus in female rats. Atrazine has been tested in male and female mice and has been determined not to be a mouse carcinogen. It is also not a *Salmonella* mutagen. Atrazine was predicted to be a rat mammary carcinogen based on the possession of 10 fragments associated with mammary cancer in other compounds. (Table 3 and Figure 1). Each of these fragments was found in three other

compounds in the learning set, all of which were mammary carcinogens (Table 3). Atrazine is therefore predicted to have a 100% chance of being a mammary carcinogen.

Fenaminosulf has been tested in male and female mice and rats and has not been observed to induce cancer in any of the four groups. Fenaminosulf however, is classified as a *Salmonella* mutagen. Fenaminosulf was predicted to be a rat mammary noncarcinogen based on seven fragments (Table 4 and Figure 2). Each of the seven fragments was found in four other chemicals in the learning set, none of which were mammary carcinogens. Fenaminosulf is therefore predicted to have a 100% chance of not being a rat mammary carcinogen.

Nithiazide has been tested in male and female mice and rats. In rats, it only induces mammary gland cancer in females. It is a noncarcinogen in male rats. Alternately, it only produces liver tumors in male mice and is a noncarcinogen in female mice. It is also a *Salmonella* mutagen. Nithiazide was predicted to be a rat mammary carcinogen based on the possession of 29 fragments. All 29 fragments were predominately found in other mammary carcinogens. However, six fragments (i.e., frags 329 to 756) also included some inactive members. As such, Nithiazide was predicted to have a 98% probability of being a rat mammary carcinogen (Table 5 and Figure 3).

Lastly, 1-phenyl-3,3-dimethyltriazene has been tested in male and female rats and was found to induce cancer of the nervous system. It has not been tested in mice. The compound is a *Salmonella* mutagen. 1-Phenyl-3,3-dimethyltriazene was predicted to be a rat nonmammary carcinogen based on the possession of eight fragments. All of the fragments except fragment

1552 were derived from other rat nonmammary carcinogens. This compound was predicted to have a 98% probability of being a nonmammary carcinogen (Table 6 and Figure 4).

Since cat-SAR predictions are based on the complete correspondence of important fragments in the model to fragments contained in the test compound, redundant reasons for activity or inactivity are observed. For example, the 10 fragments used to predict the activity of atrazine can be grouped into those just representing triazine moieties and those including part of the ethylamino and isopropylamino groups. This presents the program user with a challenge in understanding the predictions, given the redundant fragments. However, we feel that this is a strength of the cat-SAR program given the fact that it is capable of simultaneously handling thousands of fragments in the prediction and analysis processes. As such, no important features are lost due to filter mechanisms which may or may not be understood and accessible to the user. Consider that a previous SAR analysis of rat carcinogens from the CPDB based on 745 chemicals using the MultiCASE program yielded only 26 major biophores *(7)*.

Preliminary Mechanistic Analysis

Comparisons between the rat mammary carcinogen models and three other set(s) of cat-SAR models were conducted to assess the likelihood that these models were related and thus have a common underlying mechanism(s) of action. For these analyses we considered models based on rat carcinogens, female rat carcinogens, and *Salmonella* mutagens (unpublished models).

The extent of relating features between two SAR models can be taken to be indicative of the extent of mechanistic overlap between models and the underlying biological phenomena they

describe. We used the Chemical Diversity Approach (CDA) previously described by us to investigate the possible interrelationships between these models. Briefly, the CDA consists of using a random sample of 10,000 chemicals representing the "universe of chemicals". Then, using validated SAR models we predicted the activity of these chemicals. The prevalence of chemicals predicted to possess simultaneously greater than chance the ability to induce two or more toxicological effects should then provide a measure of the mechanistic relatedness of these phenomena.

The first set of CDA analyses considers the relationship between the two rat mammary carcinogen models (i.e., MC-NC and MC-NMC) and two models recently built from CPDB data for all rat carcinogens and female rat carcinogens (unpublished data). For these analyses, the level of significance was set at $p<0.001$). The rat carcinogen and female rat carcinogen models showed 43.6% greater than expected overlap (Analysis 1, Table 7). This significant overlap is expected since the learning set for the female rat carcinogen model is a subset of rat carcinogens. There was also a 72.4 % significant overlap between the rat carcinogen and rat MC-NC models (Analysis 2, Table 7). Comparison of the female rat carcinogen model to the rat MC-NC model shows a 138.0% overlap (Analysis 3, Table 7). We note that the rat MC-NC model is not a perfect subset of the female rat model since several male mammary carcinogens are included in it. However, the majority of mammary carcinogens in the MC-NC model are female mammary carcinogens. Overall, this very high degree of overlap underlines the close relationship between female carcinogens and mammary carcinogens. Taken together, these analyses suggest that the rat carcinogen, female rat carcinogen, and rat MC-NC models are all closely related.

On the other hand, the degree of overlap between the rat MC-NMC model with the rat carcinogen, female rat carcinogen, and rat MC-NC models is not high. The rat MC-NMC and rat carcinogen model had a significant but modest 19.3% overlap (Analysis 4, Table 7). The rat MC-NMC model and the female rat carcinogens model had a nonsignificant (p=0.032) overlap of only 11.4% (Analysis 4, Table 7) and the rat MC-NC and MC-NMC models also had a nonsignificant (p= 0.010) overlap of 13.6% (Analysis 5. Table 7). These last two analyses suggest that the rat MC-NMC model is significantly different than the rat MC-NC and female rat carcinogen models.

The last set of analyses considered the relationships between the carcinogen models and *Salmonella* mutagenicity. Analyses 7, 8, and 9 (Table 7) indicate a strong and expected overlap between mutagenicity and carcinogenicity. Interestingly however, there was no overlap observed between the rat MC-NMC and *Salmonella* models (Analysis 10, Table 7) (p=0.961). Since the rat MC-NMC model contained carcinogens in both its "active" (i.e., mammary carcinogens) and "inactive" (i.e., carcinogens at other sites than the mammary gland) categories, the model was standardized for carcinogens. In other words, the MC-NMC is based on mechanistic attributes that describe how carcinogens may act as breast carcinogens—not how chemicals are carcinogens.

Although the MC-NMC model was standardized for carcinogens, it also took into account a basic mechanism of carcinogenicity, that being mutagenicity. Analysis of the *Salmonella* mutagenicity of compounds in the rat MC-NC and MC-NMC models showed of the 73 mammary carcinogens with accompanying mutagenicity data from the CPDB, used for both the

MC-NC- and MC-NMC models, 61 (83.6%) were mutagens. This is consistent with findings by Gold and colleagues who reported for chemicals tested in both mice and rats, that 79% of mutagens were carcinogens and only 49% of nonmutagens were carcinogens *(4)*. Considering the MC-NC model, of the 66 noncarcinogens included in the "inactive" category that had mutagenicity data, only 14 (21%) are mutagens. This again is consistent with Gold *et al.* where they report 25% of noncarcinogens are mutagens *(4)*. On the other hand, when considering the nonmammary carcinogens used to make the "inactive" category of the MC-NMC model, of the 75 compounds with mutagenicity data, 43 (57.3%) were mutagenic. In other words, the MC-NMC model, while having carcinogens in both its "active" (i.e., mammary carcinogens) and "inactive" (i.e., nonmammary carcinogens) categories, both categories also had a high prevalence of mutagens. Thus mutagenic features were represented in both categories and therefore were not identified as being associated with activity of either category. As above, since the phenomena of mutagenesis was not considered in the modeling process, the MC-NMC describes how carcinogens may act as breast carcinogens—not how mutagens induce cancer.

Conclusions

Currently, the NTP designates 228 substances "known" or "reasonably anticipated" to pose a cancer risk *(36)*. Unfortunately, this number is based on the analysis of only a few of upwards of 70,000 chemicals manufactured and used in this country. Computational structure-activity relationships (SAR) have gained recent acceptance in the regulatory community for both human health *(37)* and ecological endpoints *(38)*. The present investigation consisted of a SAR analysis of a subset of the CPDB that included mammary carcinogens, noncarcinogens, and carcinogens at sites other than the mammary gland. Cat-SAR analysis of the MC-NC- and MC-NMC

datasets produced two sets of models with balanced sensitivity and specificity and OCP values between 70 and 82%.

Interestingly, the MC-NMC model was based on a learning set that contained carcinogens in both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories. The best of these models was able to achieve an OCP of 82%, indicating the ability to distinguish between different types of carcinogens (i.e., tissue specific), not simply between carcinogens and noncarcinogens. Moreover, based on a structural comparison between this model and a model for *Salmonella* mutagens, there was no observed relationship between the two phenomena. Likewise, in an analysis of the proportion of *Salmonella* mutagens contained in the models learning set, both the active (i.e., mammary carcinogens) and inactive (i.e., carcinogens to sites other than the mammary gland) categories had a high prevalence of mutagens. These findings suggest that the MC-NMC model is identifying structural attributes of chemicals that impart on them the ability to induce breast cancer which are separable from those generally associated with carcinogenic potential (e.g., DNA-reactivity).

By including carcinogens in the active (i.e., breast carcinogens) and inactive (i.e., non-breast carcinogens) categories of the MC-NMC, we hypothesize that we have removed a "layer" of explainable mechanisms associated with chemical carcinogenesis and have revealed another layer for studying why carcinogens target the breast (Figure 5). In other words, most SAR models, by analyzing chemicals with a known and specific biological activity, remove the majority of chemical structures and identify a subset of features associated with the given biological activity. In fact, we have demonstrated that even traditional organic chemical

categories are often removed since both the active or inactive groups being modeled contain many of these traditional features *(39)*. Likewise, since we have populated both the active and inactive categories with carcinogens, features of chemical carcinogens have been removed from the MC-NMC model allowing for features associated with how carcinogens target the breast to be identified.

Finally, the cat-SAR expert system used herein is a knowledge based one (i.e., knowledge contained in the learning set) and is not hypothesis driven. Thus, the toxicophores identified are not dependent upon previous knowledge or assumptions regarding a mechanism of action. As such, the identified attributes of breast carcinogens can be used to explore previously established or hypothesized mechanisms or more importantly, in the case of the MC-NMC model, to develop new testable hypotheses relating to the chemical induction of breast cancer.

Acknowledgments
This research was supported by the Department of Defense Breast Cancer Research Program under award number DAMD17-01-0376. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense.

Table 1. Predictive performance summary for rat mammary carcinogen – noncarcinogen (MC-NC) SAR model. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP*
ABC3/0.75							
Model 1	18021	1336	758	578	0.73(66/90)	0.78(69/88)	0.76(135/178)
Model 2	17369	1486	786	700	0.71(67/95)	0.80(72/90)	0.75(139/185)
Model 3	15547	1629	737	892	0.69(62/91)	0.76(67/88)	0.72(129/179)
Average	16979	1484	760	723	0.71	0.78	0.74(134/181
ABC3/0.90							
Model 1	18021	1016	642	374	0.82(62/76)	0.78(47/60)	0.80(109/136)
Model 2	17369	1129	617	512	0.77(56/73)	0.88(63/72)	0.82(119/145)
Model 3	15547	1311	624	687	0.83(63/76)	0.73(44/60)	0.79(107/136)
Average	16979	1152	628	524	0.81	0.80	0.81(112/139)
ABCH3/0.75							
Model 1	38797	3859	1790	2069	0.72(68/94)	0.76(68/90)	0.74(136/184)
Model 2	37636	4293	2007	2286	0.71(70/98)	0.77(75/97)	0.74(145/195)
Model 3	34407	4093	1785	2308	0.73(71/97)	0.65(62/95)	0.69(133/192)
Average	36947	4082	1861	2221	0.72	0.73	0.72
ABCH3/0.90							
Model 1	38797	2746	1434	1312	0.76(63/83)	0.78(61/78)	0.77(124/161)
Model 2	37636	2923	1392	1531	0.75(63/84)	0.78(67/86)	0.77(130/170)
Model 3	34407	2949	1372	1577	0.74(66/89)	0.71(52/73)	0.73(118/162)
Average	36947	2873	1399	2210	0.75	0.76	0.76(124/164

Footnotes:
Total Fragments: number of fragments derived from learning set.
Model Fragments: number of fragments meeting specified rules of the model.
Active Fragments: number of fragments meeting specified rules to be considered as active.
Inactive Fragments: number of fragments meeting specified rules to be considered as inactive.
Sensitivity: number of correct positive predictions / total number of positive predictions.
Specificity: number of correct negative predictions / total number of negative predictions
OCP: Observed Correct Predictions: number of correct predictions / total number of predictions.

Table 2. Predictive performance summary for rat mammary carcinogen–nonmammary carcinogen (MC-NMC) SAR model. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP*
ABC3/0.75							
Model 1	13868	1349	849	500	0.80(70/88)	0.66(53/80)	0.73(123/168)
Model 2	14461	1330	861	469	0.72(63/87)	0.72(59/82)	0.72(122/169)
Model 3	14427	1245	767	478	0.68(59/87)	0.74(64/86)	0.71(123/173)
Average	14252	1308	826	482	0.73	0.71	0.72(123/170)
ABC3/0.90							
Model 1	13868	1102	731	371	**0.83(58/70)**	**0.74(40/54)**	**0.79(98/124)**
Model 2	14461	1086	723	363	0.82(54/66)	0.72(44/64)	0.75(98/130)
Model 3	14427	847	520	327	0.82(51/62)	0.72(41/57)	0.77(92/119)
Average	14252	1308	826	482	0.82	0.73	0.77(96/124)
ABCH3/0.75							
Model 1	32235	3679	2081	1598	0.81(78/96)	0.62(55/89)	0.72(133/185)
Model 2	32374	3921	2088	1833	0.70(66/94)	0.64(59/92)	0.67(125/186)
Model 3	32627	3497	1928	1569	0.75(70/93)	0.69(65/94)	0.72(135/187)
Average	32412	3699	2032	1667	0.75	0.65	0.70
ABCH3/0.90							
Model 1	32235	2750	1642	1108	0.81(65/80)	0.76(50/66)	0.79(115/146)
Model 2	32374	2947	1637	1310	0.75(55/73)	0.69(53/77)	0.72(108/150)
Model 3	32627	2241	1170	1071	0.81(63/78)	0.70(52/74)	0.76(115/152)
Average	32412	3699	2032	1667	0.79	0.72	0.76

Footnotes: see table 1

Table 3. Fragments from the ABC 3/0.90 mammary carcinogen noncarcinogen (MC-NC) model leave-one-out validation analysis used to predict the rat mammary carcinogen atrazine.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag 3662	3	0	3	1.000	0.000
Frag 3663	3	0	3	1.000	0.000
Frag 3664	3	0	3	1.000	0.000
Frag 3665	3	0	3	1.000	0.000
Frag 3666	3	0	3	1.000	0.000
Frag 3667	3	0	3	1.000	0.000
Frag 3668	3	0	3	1.000	0.000
Frag 3670	3	0	3	1.000	0.000
Frag 3671	3	0	3	1.000	0.000
Frag 3677	3	0	3	1.000	0.000
	Probability of activity			1.00	0.00

Table 4. Fragments from the ABC 3/0.90 mammary carcinogen noncarcinogen (MC-NC) model leave-one-out validation analysis used to predict rat noncarcinogen fenaminosulf.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag 6443	0	4	4	0.000	1.000
Frag 6446	0	4	4	0.000	1.000
Frag 6447	0	4	4	0.000	1.000
Frag 6451	0	4	4	0.000	1.000
Frag 6452	0	4	4	0.000	1.000
Frag 6455	0	4	4	0.000	1.000
Frag 6461	0	4	4	0.000	1.000
	Probability of activity			0.00	1.00

Table 5. 29 Fragments from the ABC 3/90 rat nonmammary (MC-NMC) model leave-one-out validation analysis used to predict nithiazide as being a mammary carcinogen.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag328	11	1	12	0.917	0.083
Frag352	11	1	12	0.917	0.083
Frag361	12	1	13	0.923	0.077
Frag508	21	2	23	0.913	0.087
Frag746	12	1	13	0.923	0.077
Frag756	12	1	13	0.923	0.077
Frag1739	10	0	10	1.000	0.000
Frag1740	10	0	10	1.000	0.000
Frag1741	10	0	10	1.000	0.000
Frag1742	9	0	9	1.000	0.000
Frag1743	9	0	9	1.000	0.000
Frag1744	9	0	9	1.000	0.000
Frag1745	9	0	9	1.000	0.000
Frag1746	9	0	9	1.000	0.000
Frag1747	9	0	9	1.000	0.000
Frag1754	9	0	9	1.000	0.000
Frag1758	10	0	10	1.000	0.000
Frag1759	9	0	9	1.000	0.000
Frag1765	10	0	10	1.000	0.000
Frag1775	11	0	11	1.000	0.000
Frag1776	11	0	11	1.000	0.000
Frag1777	10	0	10	1.000	0.000
Frag1778	10	0	10	1.000	0.000
Frag1779	10	0	10	1.000	0.000
Frag1780	10	0	10	1.000	0.000
Frag1781	10	0	10	1.000	0.000
Frag1782	10	0	10	1.000	0.000
Frag1801	10	0	10	1.000	0.000
Frag1831	10	0	10	1.000	0.000
	Probability of activity			0.977	0.023

Table 6. 8 Fragments from the ABC 3/90 rat nonmammary (MC-NMC) model leave-one-out validation analysis used to predict 1-phenyl-3,3-dimethyltriazene as being a nonmammary carcinogen.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
Frag1552	1	11	12	0.083	0.917
Frag1557	0	5	5	0.000	1.000
Frag1558	0	5	5	0.000	1.000
Frag1559	0	5	5	0.000	1.000
Frag1561	0	5	5	0.000	1.000
Frag1572	0	6	6	0.000	1.000
Frag1576	0	6	6	0.000	1.000
Frag1577	0	5	5	0.000	1.000
	Probability of activity			0.02	0.98

Table 7. Mechanistic relationships between the cat-SAR rat mammary carcinogen, other carcinogen, and *Salmonella* mutagen models.

Analysis	*Observed*	*Expected*	*p-value*	*Δ*	*100Δ/Expected*
1. Rat + F-Rat	1166	812	<0.0001	354	43.6
2. Rat + Rat MC-NC	1410	818	<0.0001	592	72.4
3. F-Rat + Rat MC-NC	1202	505	<0.0001	697	138.0
4. Rat + Rat MC-NMC	1335	1119	<0.0001	216	19.3
5. F-Rat + Rat MC-NMC	769	690	0.032	79	11.4
6. Rat MC-NC + Rat MC-NMC	791	696	0.010	95	13.6
7. Rat + Salm	1537	1021	<0.0001	516	50.5
8. F-Rat + Salm	1648	692	<0.0001	956	138.2
9. Rat MC-NC + Salm	1595	697	<0.0001	898	128.8
10. Rat MC-NMC + Salm	933	935	0.961	-2	-0.2

Notes:
? : Observed Prevalence – Expected Prevalence
100?/Expected: 100*?/Expected Prevalence

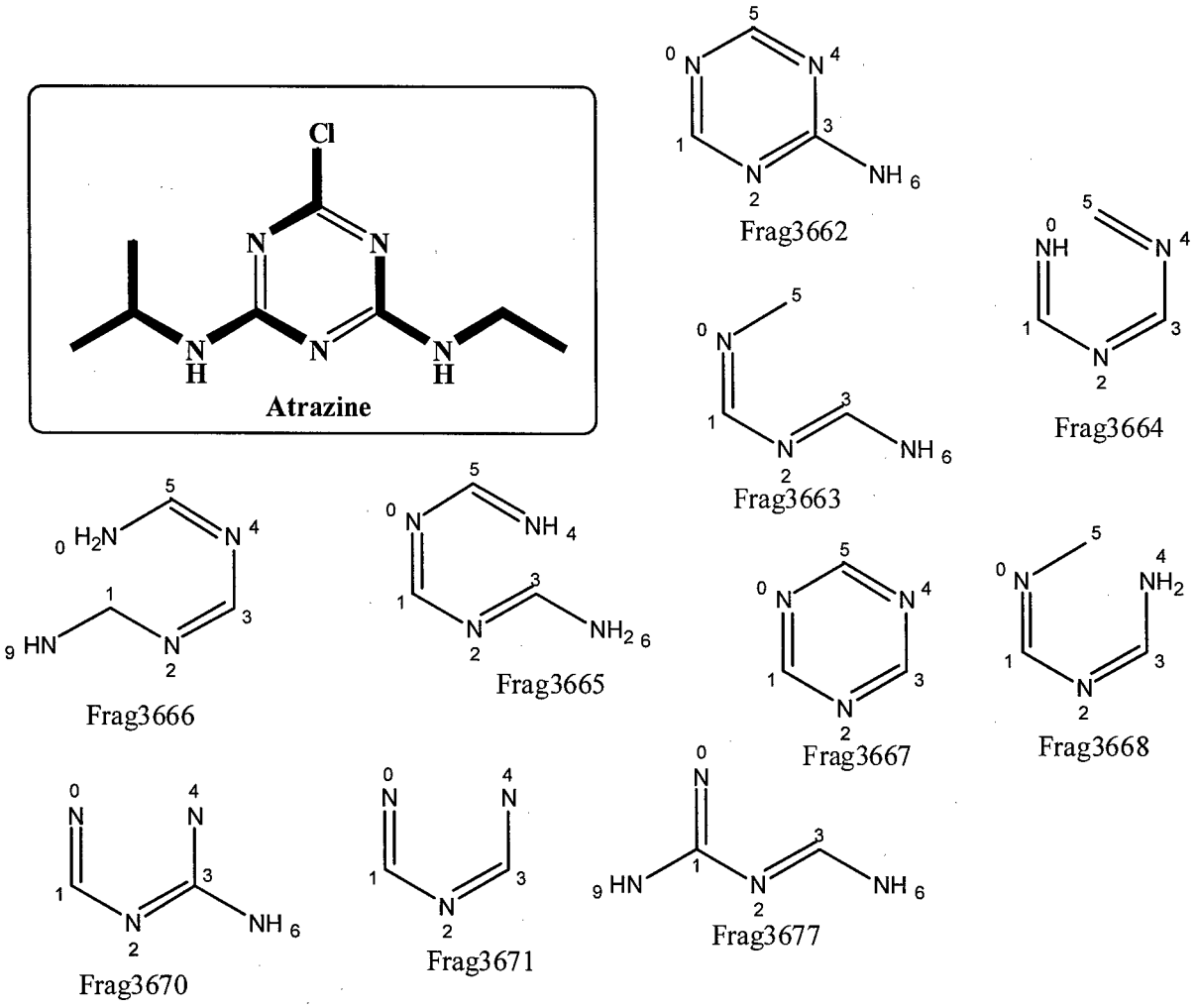


Figure 1. Illustration of the 10 significant fragments contributing to the active validation prediction of atrazine.

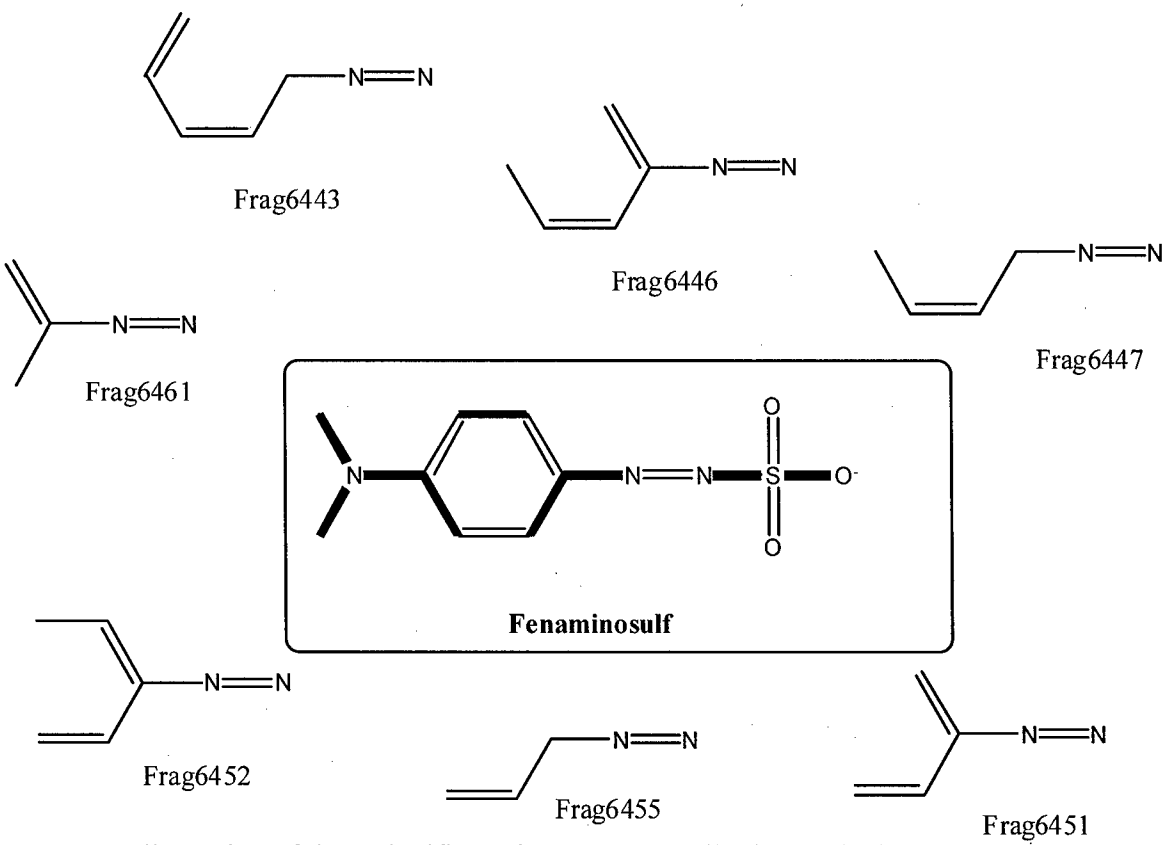


Figure 2. Illustration of the 7 significant fragments contributing to the inactive validation prediction of the fungicide fenaminosulf.

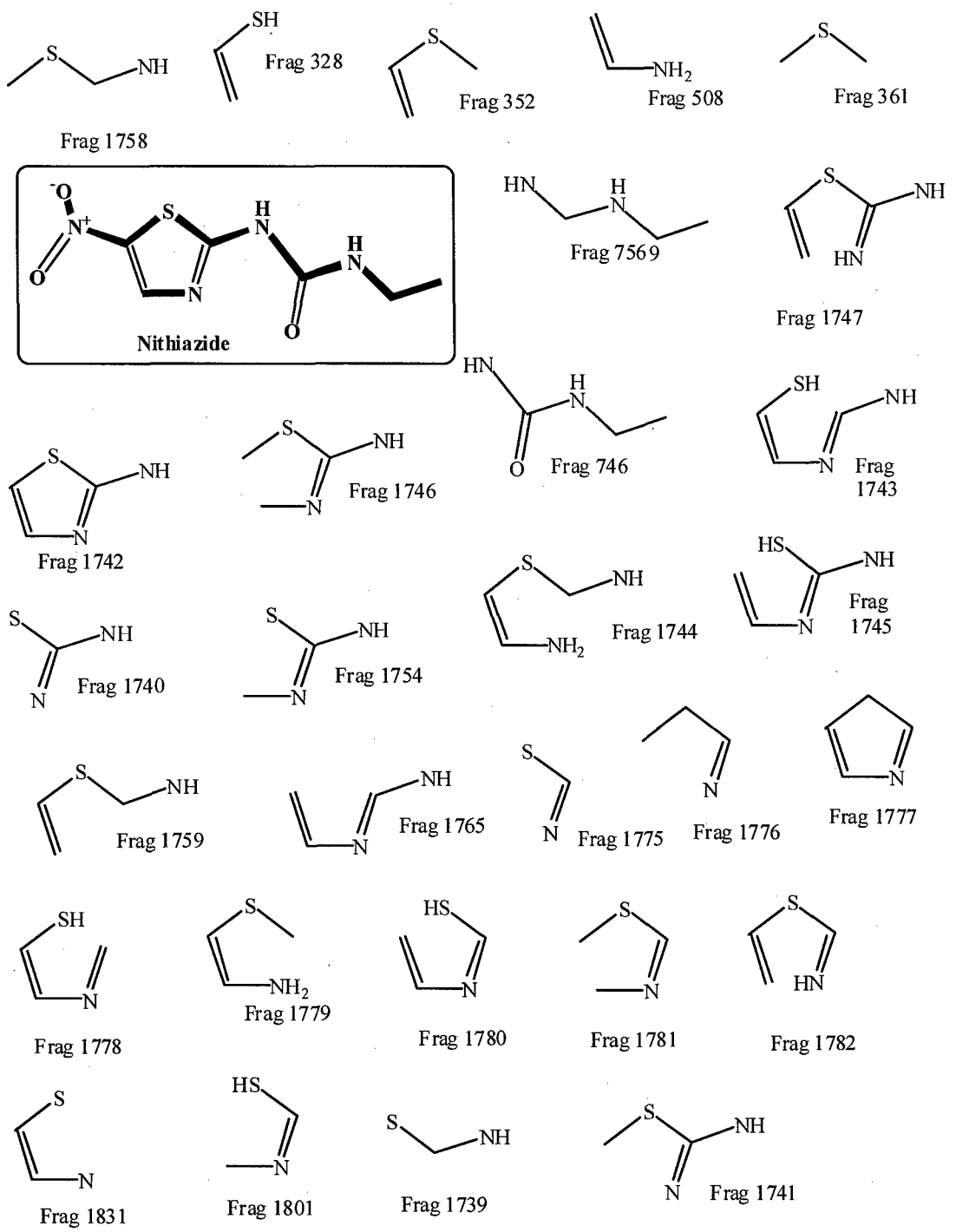


Figure 3. Illustration of the 29 significant fragments contributing to the active validation prediction of nithiazide.

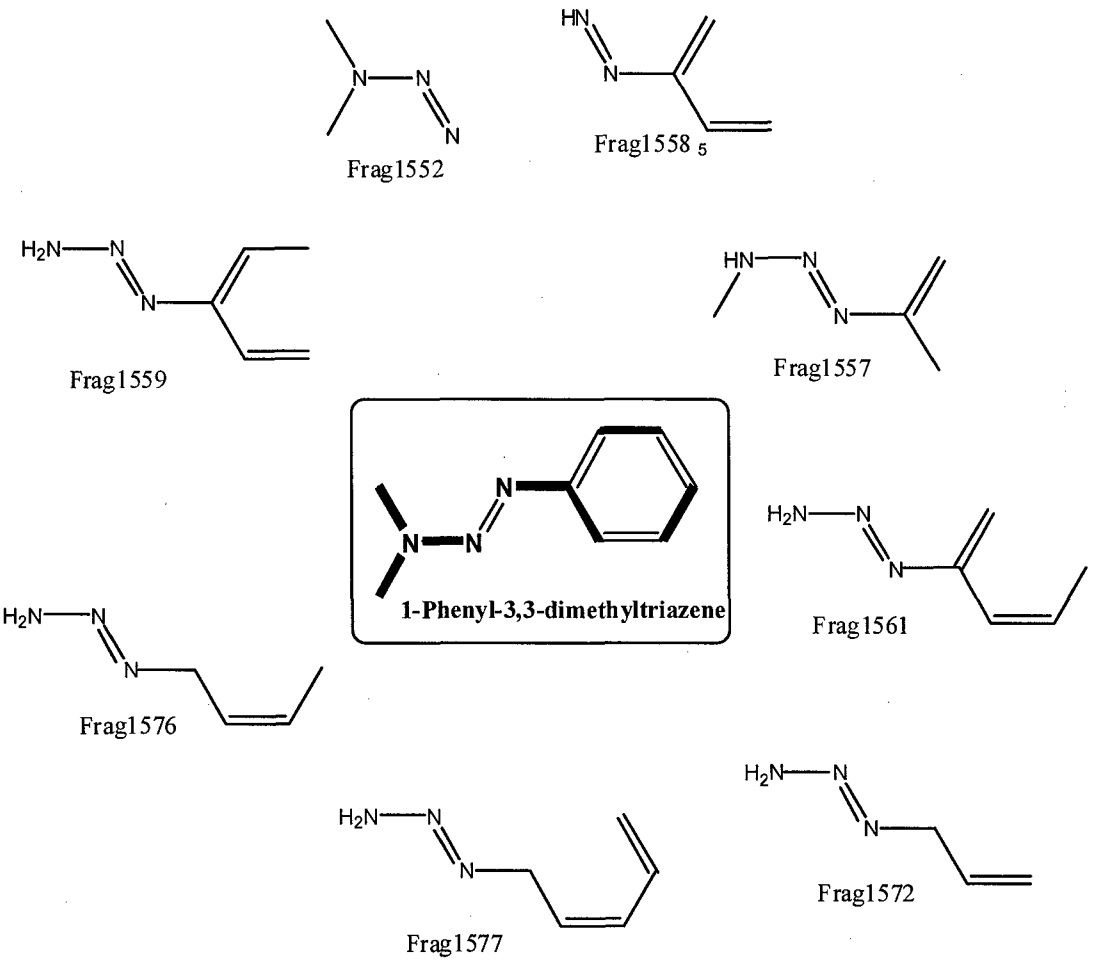


Figure 4. Illustration of the eight fragments contributing to the inactive validation prediction of 1-phenyl-3,3-dimethyltriazene.

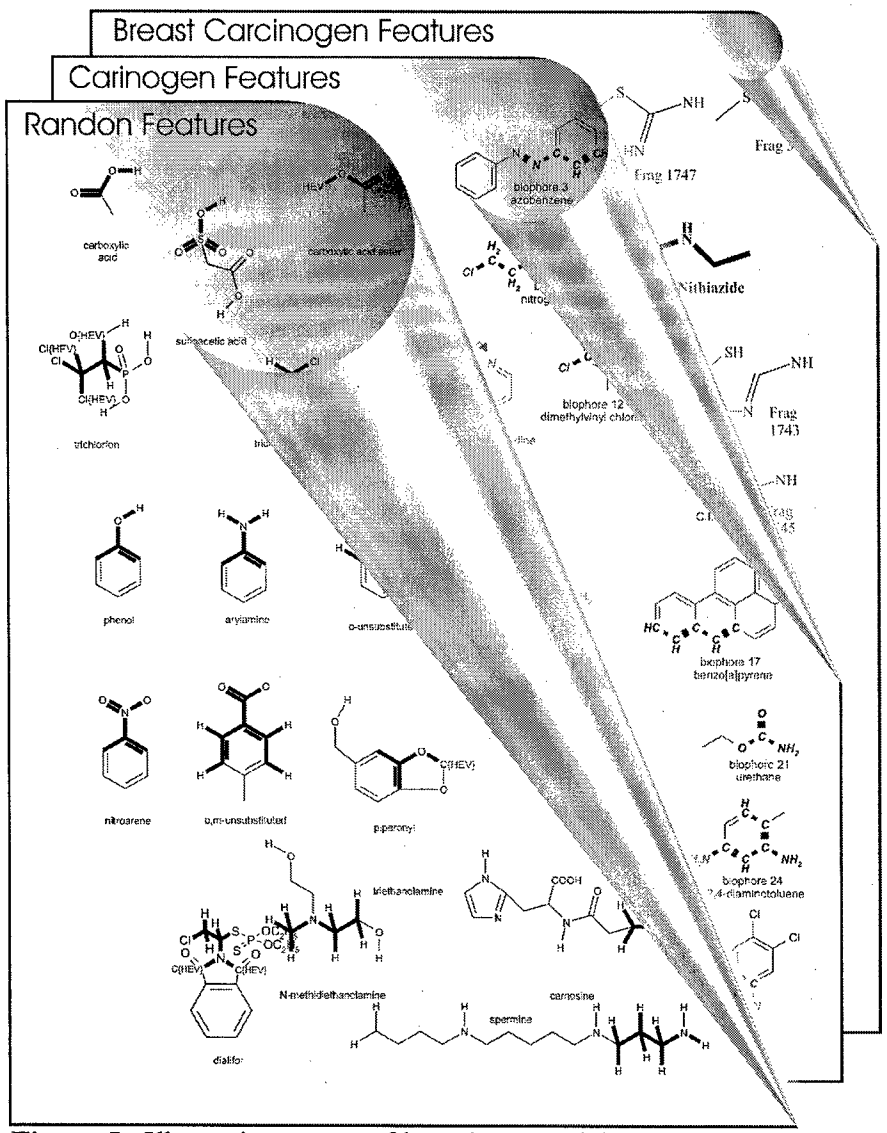


Figure 5. Illustrative nature of how SAR models can remove mechanistic layers of chemical carcinogenesis in order to study specific aspects of the process. A typical SAR model developed from categories of carcinogens and noncarcinogens removes many common chemical structures (top layer) reveals a set of features associated with carcinogenesis (middle layer). The SAR model developed from categories that both contained carcinogens (middle layer) reveals a set of features associated with breast-specificity (bottom layer).

References

(1) EPA (2004) What is the TSCA Chemical Substance Inventory? *http://www.epa.gov/oppt/newchems/invntory.htm, last accessed 10/21/04.*

(2) NIEHS (2004) About the NTP. *http://ntp-server.niehs.nih.gov/index.cfm?objectid=7201637B-BDB7-CEBA-F57E39896A08F1BB, last accessed 10/21/04.*

(3) NTP (1996) NIEHS Fact Sheet #3 The National Toxicology Program. *http://www.niehs.nih.gov/oc/factsheets/fsntp.htm, last accessed 10/21/04.*

(4) Gold, L.S., Sloan, T.H. and Ames, B.N. (1997) Overview and update of analyses of the carcinogenic potency database. In *Handbook of Carcinogenic Potency and Genotoxicity Databases* (Gold, L.S. and Zeiger, E., Eds.) pp 661-693, CRC Press, New York.

(5) CPDB (2004) Carcinogenic Potency Database. *http://potency.berkeley.edu, last accessed 10/21/04.*

(6) Cunningham, A.R., Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1998) Identification of "genotoxic" and "non-genotoxic" alerts for cancer in mice: The carcinogenic potency database. *Mutat. Res., 398*, 1-17.

(7) Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) Identification of structural features and associated mechanisms of action for carcinogens in rats. *Mutat. Res., 405*, 9-28.

(8) Gold, L.S., Sawyer, C.B., Magaw, R., Backman, G.M., deVeciana, M., Levinson, R., Hooper, N.K., Havender, W.R., Bernstein, L., Peto, R., Pike, M.C. and Ames, B.N. (1984) A carcinogenic potency database of the standardized results of animal bioassays. *Environ. Health Perspect., 58*, 9-319.

(9) Gold, L.S., deVeciana, M., Backman, G.M., Lopipero, M., Smith, M., Blumenthal, R., Levinson, R., Bernstein, L. and Ames, B.N. (1986) Chronological supplement to the carcinogenic potency database: Standardized results of animal bioassays published through December 1982. *Environ. Health Perspect., 67*, 161-200.

(10) Gold, L.S., Slone, T.H., Backman, G.M., Magaw, R., DaCosta, M., Lopipero, P., Blumenthaul, M. and Ames, B.N. (1987) Second Chronological supplement to the Carcinogenic Potency Database: Standardized results of animal bioassays published through December 1984 and by the National Toxicology Program through May 1986. *Environ. Health Perspect., 74*, 237-329.

(11) Gold, L.S., Slone, T.H., Backman, G.M., Eisenberg, S., DeCosta, M., Wong, M., Manley, N.B., Rohrbach, L. and Ames, B.N. (1990) Third chronological supplement to the Carcinogenic Potency Database: Standardized results of animal bioassays published through December 1986 and by the National Toxicology Program through June 1987. *Environ. Health Perspect., 84*, 215-286.

(12) Gold, L.S., Manley, N.B., Slone, T.H., Garfinkel, G.B., Rohrbach, L. and Ames, B.N. (1993) Fifth plot of the Carcinogenic Potency Database: Results of animal bioassays Published in the general literature through 1988 and by the National Toxicology Program through 1989. *Environ. Health Perspect., 100*, 65-168.

(13) Richard, A.M. (1998) Commercial toxicology prediction systems: a regulatory perspective. *Toxicol. Lett., 102-103*, 611-616.

(14) Richard, A.M. (1999) Application of artificial intelligence and computer-based methods to predicting chemical toxicity. *Knowl. Eng. Rev., 14*, 307-317.

(15) Miller, J.A. and Miller, E.C. (1977) Ultimate chemical carcinogens as reactive mutagenic electrophiles. In *Origins of Human Cancer* (Hiatt, H.H., Watson, J.D. and Winsten, J.A., Eds.) pp 605-627, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
(16) Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1997) A dichotomy in the lipophilicity of natural estrogens/xenoestrogens and phytoestrogens. *Environ. Health Perspect. Suppl., 105(Suppl3)*, 665-668.
(17) Rosenkranz, H.S., Cunningham, A. and Klopman, G. (1996) Identification of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens and antiestrogens. *Mutagenesis, 11*, 95-100.
(18) Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) Structural analysis of a group of phytoestrogens for the presence of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens. *Proc. Soc. Exp. Biol. Med., 217*, 288-292.
(19) Brzozowski, A.M., Pike, A.C., Dauter, Z., Hubbard, R.E., Bonn, T., Engstrom, O., Ohman, L., Greene, G.L., Gustafsson, J.A. and Carlquist, M. (1997) Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature, 389*, 753-758.
(20) Pike, A.C., Brzozowski, A.M., Hubbard, R.E., Bonn, T., Thorsell, A.-G., Engstrom, O., Ljunggren, J., Gustafsson, J.-A. and Carlquist, M. (1999) Structure of the ligand-binding domain of oestrogen receptor beta in the presense of a partial agonist and a full antagonist. *The EMBO Journal, 18*, 4608-4618.
(21) Shiau, A.K., Barsted, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A. and G.L., G. (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 297-237.
(22) Hihi, A.K. and Wahli, W. (1999) Structure and function of the estrogen receptor. In *Estrogens and Antiestrogens I* (Oettel, M. and Schillinger, E., Eds.) pp 111-126, Springer, Berlin.
(23) Davis, D.L., Bradlow, H.L., Wolff, M., Woodruff, T., Hoel, D.G. and Anton-Culver, H. (1993) Medical hypothesis: Xenoestrogens as preventable causes of breast cancer. *Environ. Health Perspect., 101*, 372-377.
(24) Davis, D.L., Axelrod, D., Bailey, L., Gaynor, M. and Sasco, A.J. (1998) Rethinking Breast Cancer Risk and the Environment: The Case for the Precutionary Principle. *Environ. Health Perspect., 106*, 523-529.
(25) Ahlborg, U.G., Lipworth, L., Titus-Ernstoff, L., Hsieh, C.C., Hanberg, A., Baron, J., Trichopoulos, D. and Adami, H.O. (1995) Organochlorine compounds in relation to breast cancer, endometrial cancer, and endometriosis: An assessment of the biological and epidemiological evidence. *Crit. Rev. Toxicol., 25*, 463-531.
(26) Ashby, J., Houthoff, E., Kennedy, S.J., Stevens, J., Bars, R., Jekat, F.W., Campbell, P., Van Miller, J., Carpanini, F.M. and Randall, G.L.P. (1997) The challenge posed by endocrine-disrupting chemicals. *Environ. Health Perspect., 105*, 164-169.
(27) Safe, S.H. (1995) Environmental and dietary estrogens and human health: Is there a problem? *Environ. Health Perspect., 103*, 346-351.
(28) Falck, F.J., Ricci, A.J. and Wolfe, M.S. (1992) Pesticides and polychlorinated biphenyl residues in human breast lipids and their relation to breast cancer. *Arch. Environ. Health, 47*, 143-146.
(29) NRC (1999) *Hormonally Active Agents in the Environment.* National Academy Press Washington, D.C.

(30) Young, J.F., Tong, W., Fang, H., Xie, Q., Pearce, B., Hashemi, R., Beger, R.D., Cheeseman, M.A., Chen, J.J., Chang, Y.-c.I. and Kodel, R.L. (2004) Building an organ-specific carcinogenic database for SAR analyses. *J. Toxicol. Environ. Health. A, 67*, 1363-1389.

(31) Gold, L.S., Manley, N.B., Slone, T.H. and Ward, J.M. (2001) Compendium of chemcal carcinogens by target organ: Results of chronic biassays in rats, mice, hamsters, dogs, and monkeys. *Toxicol. Pathol., 29*, 639-652.

(32) Dearden, J.C. (2003) *In silico* prediction of drug toxicity. *J. Comput. Aided Mol. Des., 17*, 119-127.

(33) Cunningham, A.R., Cunningham, S.L., Consoer, D.M., Moss, S.T. and Karol, M.H. (2005) Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers. *SAR QSAR Environ. Res., 16*, 273-285.

(34) Piegorsrch, W.W. and Zeiger, E. (1991) Measuring Intra-assay Agreement for the Ames *Salmonella* Assay. In *Statistical Methods in Toxicology* Statistical Methods in Toxicology ed. (Hotham, L., Ed. pp 35-41, Springer-Verlag, Heidlberg.

(35) Gold, L.S., Wright, C., Bernstein, L. and deVeciana, M. (1987) Reproducibility of results in near-replicate carcinogenesis bioassay. *J. Natl. Cancer Inst., 78*, 1149-1158.

(36) NTP "Report on Carcinogens, Tenth Edition," U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program, 2002.

(37) Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) Use of quantitative structure-activity relationships in international decision-making frameworks to predict health effects of chemical substances. *Environ. Health Perspect., 111*, 1376-1390.

(38) Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) Use of quantitative structure-activity relationships in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances. *Environ. Health Perspect., 111*, 1376-1390.

(39) Rosenkranz, H.S. and Cunningham, A.R. (2001) Chemical categories for health hazard identification: A feasibility study. *Regul. Toxicol. Pharmacol., 33*, 313-318.